IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant: William G. Thilly

Application No.: 09/503,758 Group: 1637

Filed: February 14, 2000 Examiner: T. Strzelecka

Confirmation No.: 7123

For: METHODS OF IDENTIFYING POINT MUTATIONS IN A GENOME THAT CAUSE OR ACCELERATE DISEASE

CERTIFICATE OF MAILING OR TRANSMISSION

I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as First Class Mail in an envelope addressed to Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450, or is being facsimile transmitted to the United States Patent and Trademark Office on:

2/18/05 Katie Norris

Date Signature

Katie Norris

Typed or printed name of person signing certificate

DECLARATION OF WILLIAM G. THILLY, Sc.D.
UNDER 37 C.F.R. § 1.132

Mail Stop Amendment
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

I, William G. Thilly, Sc.D., declare and state that:

1. I am the inventor on the above-referenced patent application. I am employed at Massachusetts Institute of Technology, Cambridge, Massachusetts 02139. I have been advised that Massachusetts Institute of Technology is the assignee of the entire right, title and interest of the subject application.

2. I have read the United States Patent and Trademark Office Action dated August 18, 2004, the Office Action dated January 8, 2004, the Office Action dated September 26, 2002, and the art cited by the Examiner in the Office Actions, in particular the references Kervinen et al., Artherosclerosis 105: 89-95 (1994); Margaglione et al., Stroke, 29: 399-403 (February 1998) and Paik et al., 82: 3445-3449 (1985). I have also read the patent application and the presently pending claims that were rejected in the August 18, 2004 Office Action.

3. I note that the Examiner stated in the Office Action dated August 18, 2004 that Claims have been rejected under 35 U.S.C. § 102(b) as being anticipated by Kervinen et al. as evidenced by Margaglione et al. I have read and understand the Examiner's interpretation of the Kervinen et al. and Margaglione et al. references. However, as discussed in this Declaration, the claimed invention can be distinguished from Kervinen et al. (1994) and similar published works which claimed to have found statistically significant associations between a single mutant allele in a gene and risk of a particular disease or early mortality in general.

4. Kervinen et al. (1994) and many others reviewed by Hirschhorn et al. (Genet. Med: 4(2) 45-61 (2002) attached herein as Appendix A) represent a general approach in which the frequency of each and any single allele is measured in two population samples and the frequencies are compared to discover if the absolute value of the difference is significantly different from 0.00 or if the ratio of the frequencies are significantly different from 1.00.

This approach was rooted in the widely held belief in population genetics that common diseases, including common mortal diseases, are encoded entirely or predominantly by specific single mutations in one or more genes. The examples of sickle cell anemia in African populations and cystic fibrosis among northern Europeans serve as examples of this general belief. I, however, regarded these two examples as exceptions to the general rule that inherited

diseases are encoded by multiple, tens to hundreds, of different alleles within a gene or genes, a scientific point of view now amply supported by data for nearly two thousand rare inherited human diseases (http://www.hgmd.org). In addition, recent studies have discovered a gene, *MC1R*, that encodes risk for the common diseases skin cancers by modulating the tanning response in Europeans and Asians for which some 65 putatively active alleles have been identified the summed frequencies of which total between 0.1 and 0.2 in Europeans.

As each of multiple alleles would encode a small fraction of the risk encoded by the multiple alleles, impractically large populations would need to be sampled to discern a statistically significant difference between young and aged populations for a single allele in a multi-allelic set of alleles that encoded risk for a mortal disease. Furthermore, any gene carrying alleles coding for risk of mortal disease would, as all genes, carry multiple neutral alleles that do not confer risk of mortal disease. In determining whether or not a particular gene carries alleles that encode risk the analyst does not know *a priori* the actual alleles carried by the gene in the general population. Even were the alleles known, the analyst could not specifically identify precisely which alleles conferred risk. For instance some amino acid substitutions inactivate the function of an enzyme and some do not.

5. I have devised a method that overcomes these difficulties and also reduces the size of population samples required to obtain statistically significant results. My approach has now been applied to test and negate the hypothesis that the gene *CTLA4* carried alleles conferring risk for juvenile (Type I) diabetes, a widely-held belief based inappropriately on data from a single allele of that gene.

My claimed method determines if any gene carries mutations (or alleles) in the general population that increase the risk of any common mortal disease. My method requires large samples of young and aged individuals from the same population, scanning gene segments encoding functional elements such as protein sequences and mRNA splice sites, and enumerating, or both enumerating and identifying, the set of all detectable mutations carried by any gene in both young and old populations. If a statistically significant difference in the total number of mutations exists between the young and aged groups or the total number of non-synonymous mutations or the total number of obligatory knockout mutations, i.e. the sum of stop

codons + frameshifts + mRNA splice site mutations, then the gene is identified as one that with a high degree of probability carries mutations that code for a common mortal disease. Dependent claims outline methods to extend such a finding to identify or significantly limit the number wof the many possible mortal diseases which might be caused or accelerated by the risk-conferring mutations carried b a particular identified gene.

6. I combined information from two disparate fields of research, epidemiology and mutational spectrometry, to make this invention.

From epidemiology, specifically the public health records of the United States, I organized the mortality rate data for cancers, vascular diseases and other causes of mortality from 1890 to 1997 so that the fraction of surviving persons dying from any of the diseases could be observed as a function of age (http://epidemiology.mit.edu). From these data and a self-generated mathematical model of the population in which mutations in one or more genes caused or accelerated a mortal disease applicant derived a quantitative means to estimate the expected loss of disease causing/accelerating alleles in said gene or genes as the population aged. Using pancreatic cancer as an example, it was found that between age ~50 and ~100 the fraction of the population at future risk of pancreatic cancer declined fivefold. This finding suggested that given population samples of old and young persons from the same large population, the alleles of any gene conferring risk for a mortal disease would decrease significantly between age 50 and extreme old age. I believe that prior to this work no means existed to calculate the expected fractional decrease in the alleles that encode risk for any specific mortal disease as a function of age from the public mortality records of a country.

7. From mutational spectrometry, I determined from review and organization of the existing literature that for nearly all known diseases, including mortal diseases, caused by inherited mutations in one or more genes, disease risk is encoded not by one or even two mutant alleles, but by many separate alleles distributed in a large population. This finding is demonstrated by now more than 1954 separate gene/disease relationships with an average of some 25 disease causing mutations per gene led applicant to teach the necessity of scanning a gene of interest for

a set of multiple different mutations each independently conferring disease risk. (http://www.hgmd.org).

I specifically teach the necessity of scanning all of the exons and splice sites of a gene, or as great a portion of the gene as technically practical, using the same the analytical mode for analysis of both young and aged population samples. I further teach that said scanning to discover all detectable alleles for a gene in both populations is required because it is expected that individual mutations that confer risk must be individually more rare than the sum of all such mutations.

8. I recognize that any gene will in general be found to carry a large number of mutants or alleles that do not change the molecular functionality of the gene or derived gene products. I teach that despite this fact, that in the case of a gene encoding a risk for a common mortal disease, the total number of mutations or alleles within the exons and splice sites encoding risk is large enough to permit recognition of a significant difference between young and aged populations.

The report of Kervinen et al. (1994), Margaglione et al. (1998) and similar reports must be considered in light of the above discussion about the elements of the specified method to discover if a gene carries alleles that confer risk for a mortal disease. In particular, Kervinen et al. must be considered in light of Hirschhorn et al., 2002 in which the entire class of studies represented by Kervinen et al. (1994) are found to be irreproducible and thus valueless in discovering genes that code for common diseases or common mortal diseases.

9. Kervinen et al. (1994) is one of several hundreds of studies in which high frequency mutant alleles (known as single nucleotide polymorphisms or SNPs) distributed in the general population have been tested to discover if there is a statistically significant association as indicated by a decreased frequency among the aged or an increased frequency in sample cohorts with a particular disease relative to a sample cohort of young adults drawn from the same general population. It is a matter of public record that the search based on SNPs for genes carrying alleles for common disease has failed to produce a single valid discovery. (Wall Street Journal , 14 January 2005)

Kervinen et al. (1994) specifically did not scan the gene of interest for the set of all mutations to discover if there were a significant decrease in all alleles, in all non-synonymous alleles or in all obligatory knockout alleles in aged persons as in the claimed method.

Kervinen et al. (1994) claim that their "findings strongly suggest that the presence of these potential genetic risk factors for CHD (coronary heart disease).....decreases the probability of an individual reaching an extreme old age." However, I respectfully submit that Kervinen et al. (1994) did not perform an appropriate statistical analysis and, like nearly all others who have published findings based on single allele comparisons, convinced themselves that they had observed significant age-specific allelic decline when they had not.

10. The following is a description of a standard statistical means by which allelic frequencies may be compared between any two populations. This statistical statement is then applied to the data of Kervinen et al. I have also applied the same statistical analysis to the specific teachings of this application in which some of the alleles within the exons and splice sites a particular gene encode risk for a common disease.

In general, the problem of comparing the frequency of alleles in a single gene in population A to population B is to discover if the differences in the allele frequencies are significantly greater than zero.

Let the frequency of all discovered alleles for a given gene in population A be a/A where "a" is the number of mutant alleles in a sample containing "A" total alleles (normal + mutant alleles).

Let the frequency of all discovered alleles for a given gene in population B be b/B where "b" is the number of mutant alleles in a sample containing "B" total alleles (normal + mutant alleles).

The statistical question is whether or not

$$X(a,A,b,B) = (a/A) - (b/B) > 0.$$

As a, A, b, and B may be treated as independent variables in which the values of A and B are defined one may straightforwardly calculate the variance of X as a function of derived variables in which the variances of the population sizes, A and B, defined by the experimenter are zero.

$\text{Variance}(X) = V(X) = a/A^2 + b/B^2$

$\text{Standard Deviation} = \text{Variance}^{1/2}$

$\text{Standard Deviation}(X) = (a/A^2 + b/B^2)^{1/2}$

Now the statistical question reduces to whether or not $X > 0$ reduces to the question of whether or not

$X = [(a/A) - (b/B)] - \text{quant}\ (a/A^2 + b/B^2)^{1/2} > 0$

"quant" is the multiplier derived from the Normal or Poisson distributions to define the degree of confidence that an observation has not occurred by chance. Typically biologists use the degree of confidence of (1-0.05) to indicate a significant difference between two measurements (such as weight of boys versus weight of girls) for which a quant = 1.65 expresses the desired confidence interval. However, the search for genes conferring mortal risk or causing a particular disease does not conform to this simple experiment. This is because there are more that 7.4 million common alleles or SNPs in the National Human Genome Database that affect human mortality. This large number of SNPs in addition to the gene Apo E ε4 allele examined by Kevinen et al. are "hidden" in experiments such as Kervinen et al. (1994) and unaccounted in their attempt to analyze their data.

11. This kind of statistical problem was appreciated and a solution developed by the statistician Bonferroni. He offered the expedient of defining the confidence interval such that there would be, for instance, a 0.05 probability that <u>any</u> of the pairwise possibilities lay outside the desired confidence interval by chance. This desired confidence interval he found to be

defined as the ratio: (desired degree of confidence) /(number of pairwise possibilities) In the case of trials of significance for single alleles as in Kervinen et al. the degree of confidence of (1- 0.05) would be represented by the confidence interval in which 0.05/(7.4 million) of the area under the normal distribution lie outside the interval limits

For this (1 - 0.05) degree of confidence from a two sided normal distribution the value of quant is somewhat greater than 5.5 for any trial involving any single allele of an estimated 7.4 million common alleles in the human genome. (This value of 7.40 million is the reported number of variant alleles from a population size of considerably fewer than one hundred people and must be regarded as an underestimate of the actual number of common allelic variants each occurring in 1% or more of the world 's populations.)

For this (1 - 0.05) degree of confidence from a two sided normal distribution the value of quant is about 4.0 for any trial involving counting all of the alleles in a population sample any single gene of an estimated 25,000 genes in the human genome, the method claimed in this application.

The difference in the quant values for the two kinds of approaches, single allele comparisons versus single gene, total allele comparisons, has major implications for their practical application. The sample sizes necessary to recognize significant differences in single allele differences are much greater for single allele studies than the gene scanning method taught by applicant.

Thus for experiments involving a single allele (or a large number of alleles including 7.4 million alleles) the test for a significant age-specific allelic decline must satisfy the proposition

$$X = [(a/A) - (b/B)] - 7.5\,(a/A^2 + b/B^2)^{1/2} > 0$$

For experiments involving a single allele out of 7.4 million common human alleles, the test for a significant age-specific allelic decline must satisfy the proposition

$$X = [(a/A) - (b/B)] - 5.5\,(a/A^2 + b/B^2)^{1/2} > 0$$

Table 3 of Kervinen et al. supplies the necessary values of a, A, b, B to test the hypothesis that the single e4 allele demonstrated a statistically significant decrease between the young and middle aged adults and nonagenarians (age >90).

a = number of e4 alleles (104) in young and middle aged

A = number of total alleles (520) in young and middle aged

b = number of e4 alleles (19) in nonagenarians

B = number of total alleles (190) in nonagenarians

$X = (a/A - b/B) = 104/520 - 19/190 = 0.2 - 0.1 = 0.1$

Now the question is whether or not

$X = 0.1 - 5.5 (104/520^2 + 19/190^2)^{1/2} > 0$?

or

$X = 0.1 - 5.5(0.0003846 + 0.0005263)^{1/2} > 0$?

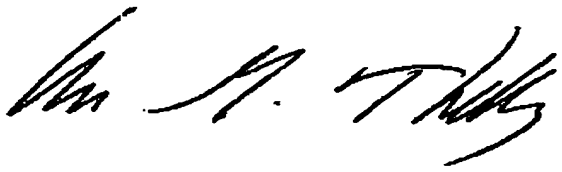
$X = 0.1 - 5.5 (0.03) = 0.1 - 0.165 = - 0.065 > 0$?

As - 0.065 is manifestly not greater than zero the claim for statistical significance of the single ε4 allele of the ApoE gene findings of Kervinen et al. (1994) cannot be sustained. As the method of Kervinen et al., used in a plethora of similar experiments reviewed by Hirschhorn et al. 2002, does not provide a means to recognize genes carrying alleles that encode risk for a mortal disease, applicant respectfully argues that it should not be used to deny applicants claims for a distinct method that can recognize such genes.

In contrast I have applied the same statistical tests to several plausible examples of risk for common mortal disease encoded by multiple alleles in the exons and splice sites of any gene as taught in both the original and amended claims. These calculations demonstrated to me that the claimed methods can determined whether or not a gene or any gene in a set of up to 25,000 genes carries risk for a common mortal disease and that said condition of risk can be discovered

by the method of scanning a gene in both young and aged population samples drawn from the same large population for all detectable mutations, all nonsynonymous mutations or all obligatory gene knockout mutations.

12. I declare that all statements made in this Declaration of my own knowledge are true and that all statements made on information and belief are believed to be true. Moreover, these statements are made with the knowledge that willful false statements and the like made by me are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.



William G. Thilly, Sc.D.

Date: 18 February 2005

review

A comprehensive review of genetic association studies

Joel N. Hirschhorn, MD, PhD[1–3], Kirk Lohmueller[1], Edward Byrne[1], and Kurt Hirschhorn, MD[4]

Most common diseases are complex genetic traits, with multiple genetic and environmental components contributing to susceptibility. It has been proposed that common genetic variants, including single nucleotide polymorphisms (SNPs), influence susceptibility to common disease. This proposal has begun to be tested in numerous studies of association between genetic variation at these common DNA polymorphisms and variation in disease susceptibility. We have performed an extensive review of such association studies. We find that over 600 positive associations between common gene variants and disease have been reported; these associations, if correct, would have tremendous importance for the prevention, prediction, and treatment of most common diseases. However, most reported associations are not robust: of the 166 putative associations which have been studied three or more times, only 6 have been consistently replicated. Interestingly, of the remaining 160 associations, well over half were observed again one or more times. We discuss the possible reasons for this irreproducibility and suggest guidelines for performing and interpreting genetic association studies. In particular, we emphasize the need for caution in drawing conclusions from a single report of an association between a genetic variant and disease susceptibility. ***Genet Med* 2002:4(2):45–61.**

Key Words: *human genetics, association studies, common disease, polymorphisms*

For most common diseases, including heart disease, diabetes, hypertension, and cancer, multiple genetic and environmental factors influence an individual's risk of being affected. This complexity contrasts with the inheritance pattern of monogenic disorders, in which the presence or absence of disease alleles usually completely predicts the presence or absence of disease (although the severity or age of onset may vary). For genetically complex diseases, risk alleles are less deterministic and more probabilistic—the presence of a high-risk allele may only mildly increase the chance of disease. Furthermore, it has been proposed that these weakly penetrant alleles may be present at high frequency (>1%) in the population.[1–3]

The widespread presence of high frequency variants in humans was first shown experimentally by Harris among others,[4] who found that many proteins have several common, heritable isoforms, thereby demonstrating that common genetic variation could lead to variation in protein structure. The widespread presence of such variation suggested that common variants might be biologically important. As Harris[4] hypothesized in 1971 (see p. 272), "The other group of alleles, though numerically much fewer, are individually much more common. They [common DNA variants] provide the basis for the great variety of enzyme . . . polymorphisms which evidently occur. These are quite possibly the underlying biochemical cause of much of the inherited diversity in the physical and physiological characteristics of individuals, and also in relative susceptibilities to various diseases and other disorders." Unfortunately, tests of this hypothesis were limited to proteins for which common functional variation could be easily assayed (primarily a few enzymes and determinants of blood group antigens).

The advent of gene cloning and sequencing substantially lowered this technical hurdle. It became possible to easily detect DNA variants in a given gene. The first genetic variants tested were usually restriction fragment length polymorphisms (RFLPs), but with the development of the polymerase chain reaction (PCR) and other improvements in technology, microsatellites, variable number tandem repeats (VNTRs), insertion/deletion polymorphisms, and single nucleotide polymorphisms (SNPs) could all be analyzed.

By determining the genotype of these variants in individuals with disease and in unaffected controls, these polymorphisms could be tested for association with susceptibility to a variety of diseases. Such studies, called "association studies," have usually used a case-control design (although family-based designs have also been used; see below). In this design, the frequencies of the alleles or genotypes at the site of interest are compared in populations of cases and controls; a higher frequency in cases is taken as evidence that the allele or genotype is associated with increased risk of disease. The usual conclusion of such studies is that the polymorphism being tested either affects risk of disease directly or is a marker for some nearby genetic variant that affects risk of disease.

From [1]Whitehead Institute/MIT Center for Genome Research, Cambridge; [2]Divisions of Genetics and Endocrinology, Children's Hospital, Boston; [3]Department of Genetics, Harvard Medical School, Boston, Massachusetts; and [4]Departments of Pediatrics and Human Genetics, Mount Sinai School of Medicine, New York, New York.

Supplementary information (full citations for references 45–663 and Supplementary Table 1) can be found at www.geneticsinmedicine.org.

Joel N. Hirschhorn, Whitehead Institute/MIT Center for Genome Research, One Kendall Square, Building 300, Cambridge, MA 02139.

Received: September 20, 2001.

Accepted: December 17, 2001.

These association studies were further facilitated by the increasingly rapid discovery of common polymorphisms in genes, accomplished by resequencing the same stretch of DNA in multiple individuals. One of the goals of the human genome project has been to identify large numbers of SNPs; indeed, the number of SNPs in public databases is now well over 1,000,000.[5] As we describe below, association studies have already identified over 600 potential associations between common genetic variants and susceptibility to common disease. As the availability of known polymorphisms skyrockets, so too will the number of reported associations. It is, therefore, critical to have a framework in place by which one can evaluate and interpret these associations.

The purpose of this publication is to list and put into perspective many of the examples of associations in the recent literature, thereby providing an interim picture of this exciting and rapidly developing field. In addition, we will examine in detail two illustrative examples: (1) the association between deep venous thrombosis and factor V Leiden, a common polymorphism in the gene encoding clotting factor V, and (2) the association between various diseases and a common polymorphism in *MTHFR*, the gene encoding methylene tetrahydrofolate reductase. Finally, we will suggest some guidelines for the analysis of association studies, because proper evaluation of these associations is critical both to understanding the genetics of common disease and to informing recent discussions regarding screening for common genetic disease.

MATERIALS AND METHODS

We performed two independent reviews of the literature from 1986 through 2000 to identify published significant associations between common diseases or dichotomous traits and common polymorphisms in or near genes (sites of genetic variation in which the minor allele frequency is at least 1%). We excluded monogenic disorders, because linkage analysis and positional cloning methods have been highly successful in identifying the alleles responsible for these diseases. Because of the large amount of prior literature, we also did not consider polymorphisms in HLA or blood group antigens, even though there are many robust associations between variation at these loci and disease. For simplicity, we have only included associations between variation at a single locus and susceptibility to disease in the entire population under study in the publication. In particular, we have not included associations between pairs of loci and susceptibility to disease nor associations between a polymorphism and susceptibility to disease in a subgroup of patients (such as smokers or those receiving hormone replacement therapy). Thereby we have explicitly ignored reports of gene–gene and gene–environment interactions, even though some of these interactions may well be of great biologic and clinical interest. Finally, we have not listed associations with substance abuse (where phenotype definition is often murky), associations between polymorphisms and variation in laboratory findings (such as serum calcium levels), or associations with other quantitative, continuous traits (as opposed to dichotomous traits). Associations were considered significant if the nominal P value was < 0.05 or if the 95% confidence intervals for relative risk excluded 1.00.

REVIEW OF THE ASSOCIATION STUDY LITERATURE

We identified 268 genes that contain polymorphisms reported to be associated with 1 of 133 common diseases or dichotomous traits. In total, these 268 genes accounted for 603 different gene-disease associations. These associations are listed in Table 1, grouped according to the trait or disease under study. As seen in Figure 1, the number of new genes associated with diseases or traits has risen more or less steadily from 1993 to 2000. The temporary drop-off in 1999 and early 2000 likely reflects an emphasis on testing newly identified polymorphisms in previously studied genes (data not shown). Examination of Table 1 also shows that many genes have been associated with several different diseases; for example, polymorphisms in *TNF*, the gene encoding tumor necrosis factor alpha, have been associated with 20 different diseases or traits, whereas variants in *ACE* (encoding angiotensin converting enzyme), *VDR* (encoding the vitamin D receptor), and *MTHFR* (encoding methylene tetrahydrofolate reductase) have each been associated with over a dozen different diseases or traits (see also supplementary Table 1). As illustrative examples, we examine in more detail two of the associations in Table 1: the association of *F5* (clotting factor V) and deep venous thrombosis, and the association between *MTHFR* and a variety of diseases.

The original report of an association between *F5* and deep venous thrombosis grew out of observations that resistance to activated protein C, a biochemically defined phenotype, was associated with markedly increased risk of deep venous thrombosis.[6] In an elegant study, the molecular basis of activated protein C resistance was shown to be a single nucleotide polymorphism in *F5* encoding an arginine to glutamine change in codon 506 (Factor V Leiden; see Bertina et al.[7]). This change occurs at one of the protein C cleavage sites, thereby preventing inactivation of factor V by activated protein C and leading to a hypercoagulable state.[8] Subsequent studies of this polymorphism have repeatedly demonstrated association with susceptibility to deep venous thrombosis, with P values often at or below 10^{-4} in individual studies (for example, Salomon et al.[9]). These studies were performed in several different populations, although the range of populations available for study is limited by the fact that Factor V Leiden is uncommon in non-Caucasian populations.[10] Thus this association is extremely robust in addition to having high biologic plausibility.

By contrast, associations involving common variation in *MTHFR* have not been as reproducible. A common thermolabile variant of methylene tetrahydrofolate reductase was first described in 1991. Thermolability of enzyme activity is inherited as a recessive trait[11] and was eventually shown to be due to homozygosity for the "T" allele at a C/T polymorphism in nucleotide 677 (causing an alanine to valine change, see Frosst et al.[12]). Unlike the rare, more severe mutations in *MTHFR*

Table 1
Associations between common polymorphisms in genes and common diseases or dichotomous traits

Disease/trait	Gene (ref)	Gene (ref)	Gene (ref)	Gene (ref)
Cancer				
Acute leukemia	CYP1A1 (45)	CYP2D6 (45, 46)	GSTM1 (45)	GSTT1 (45)
	MTHFR (20)	NAT2 (47)		
Bladder cancer	GSTM1 (48)	GSTP1 (49)	GSTT1 (50)	
Breast cancer	COMT (51)	CYP17 (52)	CYP19 (53)	CYP1A1 (54)
	CYP1B1 (55, 56)	ERBB2 (57)	ESR1 (58)	GSTM1 (59)
	HRAS (60)	HSPA8 (61)	NAT1 (62)	NAT2 (63)
	PGR (64)	SHBG (65)	SOD2 (66)	TP53 (67)
	VDR (68)			
Cervical cancer	GSTT1 (69)	MTHFR (70)	TP53 (71)	
CLL	ETS1 (72)	TNF (73)		
Colorectal cancer	ALDH2 (74)	APC (75)	CYP1A1 (76)	DIA4 (77)
	GSTM1 (78)	GSTT1 (79)	LTA (80)	MSH3 (81)
	MTHFR (18)	NAT1 (82)	NAT2 (83)	XRCC1 (84)
Endometrial cancer	CDKN1A (85)	CYP1A1 (86)	MMP1 (87)	MTHFR (86)
	TP53 (88)			
Gastric cancer	ALDH2 (74)	GSTM1 (89)	GSTT1 (90)	IL1B (91)
	MYC (92)			
Glioblastoma	PPARG (93)			
Head/neck cancer	ADH1B (94)	ALDH2 (94)	CDKN1A (95)	CYP1A1 (96)
	CYP2D6 (97)	CYP2E (98)	FCGR3A (99)	GSTM1 (100)
	GSTM3 (101)	GSTP1 (102)	GSTT1 (101)	LTA (103)
	MYCL1 (104)	NAT1 (48)	NAT2 (102, 105)	TP53 (106)
Hodgkin's lymphoma	HSPA8 (61)	TNF (61)		
Liver cancer	CYP2D6 (107)	CYP2E (108)	EPHX1 (109)	
Lung cancer	ALDH2 (74)	CDKN1A (110)	CYP1A1 (111)	CYP1B1 (55)
	CYP2A6 (112)	CYP2E (113)	DIA4 (114)	DIA4 (115)
	EPHX1 (116)	GPX1 (117)	GSTM1 (118, 119)	HRAS (120)
	LTA (121)	MGMT (122)	MPO (123)	NAT1 (124, 125)
	NAT2 (126)	TF (127)	TP53 (128)	
Melanoma	HRAS (129)	MC1R (130)	XRCC3 (131)	
Non-Hodgkin's lymphoma	EPHX1	ETS1 (132)	PGR	
Oral leukoplakia	GSTM1 (133, 134)	GSTT1 (133, 134)		
Oligoastrocytoma	ERCC1 (99)			
Ovarian cancer	HRAS (135)	TP53 (136)		
Prostate cancer	AR (137, 138)	CYP17 (139, 140)	CYP1A1 (141)	CYP1B1 (142)
	CYP3A4 (143)	ELAC2 (144)	GSTP1 (49)	SRD5A2 (145)
	VDR (146)			
Renal cell cancer	CYP1A1 (147)	GSTT1 (148)		
Testicular cancer	GSTP1 (49)			
Cardiovascular disease				
CAD/MI	ACE (149)	ADRB3 (150)	AGTR1 (151)	APOA1 (152)
	APOB (153)	APOE (154)	CD14 (155)	CYBA (156)
	F13A1 (157)	F2 (158)	F5 (159)	F7 (160)
	FGB (161)	GP1BA (162)	GSTM1 (163)	HTR2A (164)
	IRS1 (165)	ITGA2 (166)	ITGB3 (167)	LPL (168)
	MMP3 (169)	MTHFR (13, 14)	NOS3 (170, 171)	NPPA (172)
	PLAT (173)	PON1 (174)	PON2 (175)	PPARG (176)
	SELE (177)	SELP (178)	SERPINA8 (179, 180)	SERPINE1 (181)
	TGFB1 (182)	THBD (183)	WRN (184)	
DVT	F13A1 (185)	F2 (186)	F3 (187)	F5 (7)
	MTHFR (19)	PLAT (188)	PON1 (189)	
Dilated cardiomyopathy	ACE (190)	EDNRA (191)	PLA2G7 (170)	SOD2 (192)
HTN	ACE (193)	ADD1 (194)	AGTR1 (195)	CYP11B2 (196)
	DIA4 (197, 198)	DRD1 (199)	GCK (200)	GNAS1 (201)
	GNB3 (202)	GYS1 (203)	HSD11B2 (204)	INSR (205)
	MTHFR (206)	NPPA (172)	REN (207)	SAH (208)
	SCNN1B (209)	SERPINA8 (210)	TGFB1 (182)	TH (211)
Survival post-CHF	ADRB2 (212)	AMPD1 (213)		
Dermatology				
Acne	MUC1 (214)			
Contact dermatitis	NAT2 (215)			
Eczema	CMA1 (216)			
Psoriasis	C4A (217)	CDSN (218)	LTA (219)	OTF3 (220)
	SERPINA8 (219)	TAP1 (221)	TNF (222, 223)	VDR (224)

— Continued

Table 1
(Continued)

Disease/trait	Gene (ref)	Gene (ref)	Gene (ref)	Gene (ref)
Endocrinology				
Addison's disease	CTLA4 (225)			
Gestational DM	INSR (226)			
Graves' disease	CTLA4 (227)	IFNG (228)	IL4 (229)	TAP1 (230)
	THRB (231)	TRHR (232)	VDR (233)	
Hyperparathyroidism	VDR (234)			
Male infertility	AR (235)	LHB (236)		
Obesity	ABCC8 (237)	ADRB2 (238)	ADRB3 (239)	APOB (240)
	APOD (241)	GNB3 (242)	LDLR (243)	LEP (244)
	LIPE (245)	NMB (246)	NPY5R (247)	PPARG (248)
	TNF (249)			
Osteoporosis/fracture	COL1A1 (250)	TGFB1 (251)	VDR (252)	
PCOS	CYP11A (253)	CYP17 (254)	FSHB (255)	FST (256)
	INS (257)	LHB (258)		
Short stature	DRD2 (259)	VDR (260, 261)		
Type 1 diabetes	BCL2 (262)	C4A (263)	CCR2 (264)	CD3D (265)
	CD4 (265)	CTLA4 (266)	GCK (267)	ICAM1 (268, 269)
	IFNG (270)	IGHV2-5 (271)	IL6 (272)	INS (273)
	LTA (274)	NEUROD1 (275)	PSMB8 (276)	VDR (277)
	WFS1 (278)			
Type 2 diabetes	ABCC8 (279)	ACE (280)	ADRB2 (281, 282)	CD4 (283)
	FRDA (284)	GCGR (285, 286)	GCK (287, 288)	GYS1 (289)
	HFE (290)	INS (291)	INSR (292, 293)	IPF1 (294)
	IRS1 (295)	KCNJ11 (296)	PCSK2 (297)	PPARG (37)
	PPP1R3 (298)	RRAD (299)	SLC2A1 (300)	SLC2A2 (301)
	TCF1 (302)	UCP3 (303)		
Gastroenterology				
Celiac disease	CTLA4 (304)	TNF (305)		
Cholelithiasis	APOB (306)	CETP (307)		
IBD	BDKRB1 (308)	F5 (309)	IL10 (310)	IL1RN (311)
	MLH1 (312)	MTHFR (313)	MUC3A (314)	TNF (315)
	VDR (316)			
	IL1RN (317)			
Pancreatitis				
Primary biliary cirrhosis	CTLA4 (318)	VDR (319)		
Infectious disease				
Cerebral malaria	CD36 (320)	ICAM1 (321)	NOS2A (322)	TNF (323)
HIV infection/AIDS	CCR2 (324)	CCR5 (325, 326)	CX3CR1 (327)	MBL2 (328)
	SDF1 (329)	SLC11A1 (330)		
Leishmaniasis	TNF (331)			
Leprosy	TNF (332)	VDR (333)		
Meningococcal disease	FCGR2A (334)	SERPINE1 (335)	TNF (336)	
Parasitic infections	ADRB2 (337)	NOS2A (338)		
RSV bronchiolitis	IL8 (339)			
Severe sepsis	IL1RN (340)			
Trachoma	IL10 (341)	TNF (342)		
Tuberculosis	SLC11A1 (343)			
Viral hepatitis	MBL2 (344)	TNF (345)		
Miscellaneous				
Athletic endurance	ACE (346)			
Benzene toxicity	DIA4 (347)			
Fair skin, red hair	MC1R (348)			
High altitude HTN	ACE (349)			
Lead poisoning	ALAD (350)			
Longevity	ACE (351)	APOA1 (352)	APOB (353)	APOE (354)
	SERPINE1 (355)			
Macular degeneration	APOE (356)	EPHX1 (357)	SOD2 (357)	
Tobacco use	DRD2 (358)	SLC6A3 (359)		
Trichloroethylene toxicity	GSTM1 (360)	GSTT1 (360)		
Neonatal disease				
Cleft lip/palate	BCL3 (361)	MSX1 (362)	RARA (363)	TGFA (364)
	TGFB2 (365)	TGFB3 (362)		
Neural tube defect	MTHFR (16, 17)	MTR (366)	T (367)	
Pyloric stenosis	NOS1 (368)			
RDS	SFTPA1 (369, 370)			

— Continued

Table 1
(Continued)

Disease/trait	Gene (ref)	Gene (ref)	Gene (ref)	Gene (ref)
Neurology				
Absence seizures	GABRB3 (371)	OPRM1 (372)	SLC6A3 (373)	
Alzheimer's disease	A2M (374, 375)	ACE (376)	APBB1 (377)	APOA4 (378)
	APOC1 (379)	APOC2 (380)	APOE (381)	BCHE (382)
	BLMH (383)	IL1A (386)	CTSD (384)	HTR6 (385)
	LRP1 (387)	NOS3 (388)	PSEN1 (389)	SERPINA3 (390)
	SLC6A4 (391)	TF (392)	TFCP2 (393)	TGFB1 (394)
	TNFRFS6 (395)	VLDLR (396)		
Creutzfeldt-Jakob disease	PRNP (397)			
Epilepsy	CHRNA4 (398)			
Guillian-barré syndrome	TNF (399)			
Head injury outcome	APOE (400)			
Hydrocephalus	APOE (401)			
Intracranial aneurysms	ACE (402)	ENG (403)	MMP9 (404)	
Ischemic stroke	ACE (405)	APOE (406)	CYBA (407)	ENG (408)
	F13A1 (409)	F2 (410)	FGB (411)	GP1BA (162)
	ITGA2 (412)	MTHFR (413, 414)	NOS3 (415)	NPPA (416)
	PLA2G7 (417)	PON1 (418)		
Migraine headache	DBH (419)	MTHFR (420)	SLC6A4 (421)	
Multiple sclerosis	CTLA4 (422)	IL1RN (423)	MBL2 (424)	PTPRC (425)
Myasthenia gravis	FCGR2A (426)	IL1B (427)	TNF (428)	
Otosclerosis	COL1A1 (429)			
Parkinson's disease	A2M (430)	ADH4 (431)	CCK (432)	COMT (433)
	CYP1A1 (434)	CYP2D6 (435)	DLST (436)	DRD2 (437)
	EPHX1 (438)	GSTP1 (439)	MAOA (440)	MAOB (441)
	MAPT (442)	NAT2 (443)	NOS3 (444)	SERPINA3 (445)
	SERPINA3 (445)	SLC6A3 (446)	SLC6A4 (447)	SNCA (448)
	UCHL1 (449)			
Obstetric disease				
Endometriosis	ESR1 (450)			
Fetal loss	ACP1 (451)	CTLA4 (452)	EPHX1 (453)	F2 (454)
	F5 (455)	MTHFR (456)		
Preeclampsia	AGTR1 (457)	F2 (458)	F5 (459)	LPL (460)
	MTHFR (461)	NOS3 (462)	SERPINE1 (463)	TNF (464)
Pharmacogenetics				
Albuterol response	ADRB2 (465)			
Antidepressant response	GNB3 (466)			
Aspirin response	ITGB3 (467)			
Azathioprine toxicity	TPMT (468)			
Beta-blocker response	GNAS1 (201)			
Clozapine response	DRD3 (469)	HSPA1A (470)	HSPA2 (470)	HTR2A (471)
	HTR2C (472)	HTR6 (473)	TNF (474)	
Drug-induced tardive dyskinesia	CYP2D6 (475, 476)	DRD2 (477)	DRD3 (478)	HTR2C (479)
	SOD2 (480)			
Fluvastatin response	APOB (481)			
Fluvoxamine response	SLC6A4 (482)			
Irinotecan toxicity	UGT1A1 (483)			
Leukotriene Inhibitor response	ALOX5 (484)			
Lithium response	IMPA1 (485)			
Menadione-associated urolithiasis	DIA4 (486)			
Omeprazole response	CYP2C19 (487, 488)			
Pravastatin response	CETP (489)	MMP3 (490)		
Tacrine response	APOE (491)			
Tricylic antidepressant response	CYP2D6 (492)			
Warfarin response	CYP2C9 (493)			

—Continued

which cause homocystinuria, the variant was not associated with neurologic deficits. However, thermolability of enzyme activity was observed to be associated with altered homocysteine levels and risk of coronary artery disease,[11] findings that were confirmed in at least one subsequent study that looked at nucleotide 677 (see Gallagher et al. and Kluijtmans et al.[13,14]). Folate metabolism and homocysteine levels are connected with several clinical disorders, including coronary artery disease, deep venous thrombosis, neural tube defects, and cancer (see Gailey and Gregory[15] for review); the thermolabile variant has been associated in different studies with increased risk of each of these diseases.[13,14,16–20] However, despite the biologic plausibility of these associations, none have been reproducibly observed across many studies (for example, Ma et al.[21–23]).

If all of the associations listed in Table 1 could be replicated as consistently as factor V Leiden and deep venous thrombosis,

Table 1
(Continued)

Disease/trait	Gene (ref)	Gene (ref)	Gene (ref)	Gene (ref)
Psychiatry				
Anorexia	HTR2A (494)			
ADHD	COMT (495)	DRD4 (496)	DRD5 (497)	SLC6A3 (498)
	HTR2A (499)	SNAP25 (500)		
Autism	ADA (501)	EN2 (502)	FMR1 (503)	
Bipolar disorder	APOE (504)	ATP1A3 (505)	COMT (506)	DDC (507)
	DRD3 (508)	GABRA5 (509)	HTR5A (510)	HTR6 (511)
	MAOA (512)	MAOB (513)	PLA2G1B (514)	PLCG1 (515)
	SERPINA8 (516)	SLC6A4 (517)	TPH (518)	
Compulsive gambling	DRD2 (519)	DRD4 (520)		
Depression	ACE (521)	COMT (522)	DRD3 (523)	DRD4 (524)
	GNB3 (466)	HTR5A (510)	SLC6A4 (525)	TPH (526)
OCD	DRD4 (527)	HTR1B (528)	HTR2A (529)	SLC6A4 (530)
Panic disorder	ADORA2A (531)	CCK (532)		
Schizophrenia	APOE (533)	CCK (534)	CCKBR (535)	COMT (536)
	DRD2 (537)	DRD3 (538)	DRD4 (539)	DRD5 (540)
	GNAL (541)	HMBS (542)	HRH2 (543)	HTR2A (544)
	HTR5A (510)	HTR6 (545)	KCNN3 (546)	NTF3 (547)
	OPRS1 (548)	PLA2G4A (549)	PLA2G7 (550)	YWHAH (551)
Pulmonary disease				
Asthma/atopy	ACE (552)	ADRB2 (553)	CCR5 (554)	CFTR (555)
	GSTP1 (556)	HNMT (557)	IL10 (558)	IL13 (559)
	IL4 (560)	IL4R (561)	IL9R (562)	LTA (563)
	MS4A1 (564)	NOS1 (565)	NOS3 (566)	PLA2G7 (567)
	SCYA5 (568)	SERPINA8 (569)	TAP1 (570)	TAP2 (571)
	TBXA2R (572)	TNF (563)	UGB (573)	
COPD/emphysema	CFTR (574)	EPHX1 (575)	GC (576)	GSTP1 (577)
	SERPINA1 (578)	SERPINA3 (579)	TNF (580)	
Pneumoconiosis	TNF (581)			
Pulmonary fibrosis	TGFB1 (582)			
Pulmonary embolism	FGA (583)			
Sarcoidosis	ACE (584)	CCR2 (585)	CCR5 (586)	SLC11A1 (587)
	VDR (588)			
Renal/urologic disease				
IgA nephropathy	TRA@ (589)			
Nephrotic syndrome	SERPINA1 (590)			
Renal failure	BDKRB1 (591)	DCP1 (592)	HSD11B2 (593)	KLKB1 (594)
	NOS3 (595)	SERPINA8 (592)		
Urolithiasis	DIA4 (486)			
Rheumatology				
Behcet's disease	ICAM1 (596)			
Intervertebral disc disease	COL9A2 (597)			
Juvenile chronic arthritis	IL6 (598)	TAP2 (599)		
JRA	SLC11A1 (600)			
Osteoarthritis	COL2A1 (601)	VDR (602)		
Rheumatoid arthritis	CRH (603, 604)	ESR1 (605)	HSPA1A (606)	IFNG (607)
	SLC11A1 (608)	TAP2 (609)	TRD@ (610)	XRCC3 (611, 612)
Sjogren's syndrome	GSTM1 (613)			
SLE	ACE (614)	ADPRT (615)	BCL2 (262)	C4A (427)
	C4B (616)	CTLA4 (617)	CYP2D6 (618)	FCGR2A (619)
	HSPA2 (620)	IGHV3-30-5 (621)	IL10 (622)	MBL2 (623)
	TNF (624)	VDR (625)		
Wegener's granulomatosis	CTLA4 (626)	PRTN3 (627)		

For each disease or trait, the number(s) in parentheses identifies the first reference(s) reporting a significant association with a polymorphism in the gene indicated by its official symbol. Citations can be found at *www.geneticsinmedicine.org*. Full gene names and OMIM numbers are listed in Table 4. CLL, chronic lymphocytic leukemia; CAD/MI, coronary artery disease/myocardial infarction; HTN, hypertension; CHF, congestive heart failure; DM, diabetes mellitus; PCOS, polycystic ovary syndrome; IBD, inflammatory bowel disease; RDS, respiratory distress syndrome; ADHD, attention deficit hyperactivity disorder; OCD, obsessive compulsive disorder; COPD, chronic obstructive pulmonary disease; JRA, juvenile rheumatoid arthritis; SLE, systemic lupus erythematosius; RSV, respiratory syncytial virus; DVT, deep vein thrombosis; IgA, immunoglobulin A.

this list would represent a significant understanding of the etiologies of most of the major human diseases. However, genetic associations more often behave like those seen with *MTHFR*: they are not consistently reproducible. To determine what fraction of the associations in Table 1 were robust, we first identified those associations for which an assessment of reproducibility could be made. These 166 associations (those for which we could find and review at least three separate publications) are listed in Table 2. Where more than one polymorphism in a gene was studied, the polymorphisms were treated

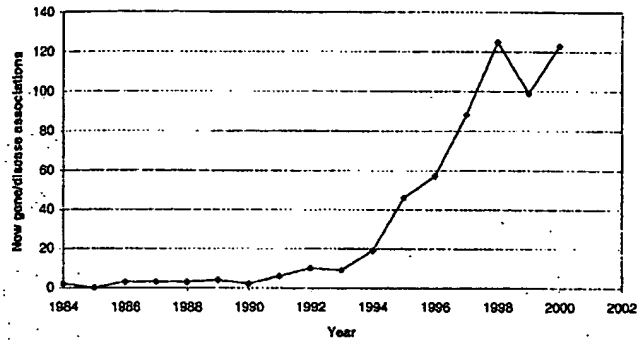


Fig. 1 The number of new, previously unreported, significant associations between diseases or dichotomous traits and genes is plotted for each year from 1984 through 2000. The graph does not include new associations between a disease or trait and polymorphisms in a gene for which other polymorphisms had previously been significantly associated with that disease or trait.

separately. Although a significant effort was made to be complete, there are undoubtedly some well-studied associations that are not listed in Table 2. Nevertheless, we believe that this list is a reasonably accurate representation of the state of published association studies between polymorphisms and common genetic disease.

We reviewed the 166 associations in Table 2 to determine whether other studies of the same polymorphism and disease also reached statistical significance. Only six associations were reproduced at a high level of consistency (statistical significance was achieved in 75% or more of all identified studies). These six associations are listed in Table 3. The possibility of publication bias and consequent omission of "negative" studies means that six is actually an upper limit for the number of consistently reproducible associations. Of the associations in Table 3, the most reproducible was the association of ApoE4 and Alzheimer's disease, for which dozens of reports reach statistical significance. It should be noted, however, that the association is most robust in Caucasians (all identified reports achieved statistical significance); for other ethnic groups (Africans, African-Americans, and Hispanics), the association is sometimes more difficult to demonstrate.[24–26]

What could be the cause of the irreproducibility that characterizes the vast majority of association studies? One possibility is that the original observations represent statistical fluctuations (type I error). If this were the case, one would predict that only 5% of subsequent studies would also reach statistical significance with $P < 0.05$, and most associations would never be observed again. However, of the 166 associations listed in Table 2, at least 97 were observed again, many of them multiple times. Thus in the absence of a massive publication bias (selective publication of positive results with numerous negative studies remaining unpublished), statistical fluctuation is unlikely to explain all of the initial positive reports in Table 2.

Other possible causes of false-positive association studies have been previously identified and include ethnic admixture resulting in population stratification, variable linkage disequilibrium between the polymorphism being studied and the true causal variant, and population-specific gene–gene or gene–environment interactions.[27–30] Each of these issues is addressed briefly in turn below, and possible remedies are offered. Finally, we examine the possibility that weak genetic effects combined with underpowered studies lead to significant numbers of falsely negative reports.

POPULATION STRATIFICATION

Most association studies have a case-control study design, in which allele or genotype frequencies in patients are compared with frequencies in an unaffected control population (Fig. 2a). This study design is subject to population stratification due to ethnic admixture, which occurs when the cases and controls are unintentionally drawn from two or more ethnic groups or subgroups. If one of these subgroups has a higher disease prevalence than the others, stratification occurs, because that subgroup will be overrepresented in the cases and underrepresented in the controls. Any polymorphism that genetically marks the high-risk subgroup (i.e., is found by chance at a higher frequency in that subgroup), therefore, will appear to be associated with disease (Fig. 2b) and will likely be a false positive. Interestingly, the frequencies of several of the alleles in Table 2 vary substantially between populations, consistent with the possibility of false associations due to ethnic admixture. It should be noted that well-defined subgroups are not necessary to observe stratification; stratification can also occur in a single admixed population where the individuals have varying degrees of genetic contributions from two or more ethnic groups. Even apparently homogeneous, isolated populations (such as Iceland) are in theory susceptible to admixture if there have been multiple distinct waves of migration from different source populations (e.g., Celtic and Norse, in the case of Iceland).

What steps can be taken to prevent false-positive associations due to population stratification? Currently, two solutions can be attempted. First, one can use family-based studies such as the transmission disequilibrium test.[31] This method, abbreviated TDT, requires affected offspring and their parents to test an allele for association with disease; the frequency with which heterozygous parents transmit that allele to offspring is then determined. This frequency is compared with the Mendelian expectation of 50:50 transmission of the allele. TDT (like other family-based methods) is immune to false-positives from ethnic admixture.[31] Disadvantages of the TDT are that family-based samples are often difficult to collect and that 50% more genotyping is required than in case-control studies to achieve similar power (the exact loss of power depends on the underlying genetic model). Another possibility is to study multiple case-control populations, each from different ethnic groups, and require that an association be seen in each population. Finally, an approach to detect and correct for stratification has been proposed: by typing several dozen random markers, one can empirically determine the degree of stratification in a case control study.[32–34] If significant stratification is detected, one can use these markers to more carefully match cases and controls to remove the effects of stratification.[35] There is some debate as to whether stratification is a significant problem;

Table 2
Disease-polymorphism associations for which at least three studies were identified

Disease/trait	Gene	Polymorphism	Risk allele/genotype	Frequency	Reference
Cancer					
Bladder cancer	GSTM1	null (gene deletion)	null/null	0.48–0.60	628, 629
Bladder cancer	GSTT1	null (gene deletion)	null/null	0.15	50
Bladder cancer	NAT2	857G/A = BamHI	A = M3 = slow acetylator	0.06	630
Breast cancer	CYP17	−34T/C = MspAI	T/C and C/C	0.55	52
Breast cancer	CYP1A1	3′ C/T (MspI)	site present/site present = C/C	0.04	54
Breast cancer	GSTM1	null (gene deletion)	null/null	0.46	59
Cervical cancer	TP53	Pro72Arg	Arg	0.66	71
Colorectal cancer	GSTM1	null (gene deletion)	null/null	0.42	78
Colorectal cancer	NAT2	590G/A = TaqI	G/G (no *6 alleles)	0.51	83
Head/neck cancer	CYP1A1	Ile462Val	Ile/Val and Val/Val	0.08	96
Head/neck cancer	CYP1A1	3′ C/T (MspI)	site present = C	0.23	631
Head/neck cancer	CYP2E	5′ RsaI site	site present/site present	0.56	98
Head/neck cancer	GSTM1	null (gene deletion)	null/null	0.48	100
Head/neck cancer	GSTM3	A/B (MnlI)	B/B	0.06	101
Head/neck cancer	GSTP1	Ile104Val = A313G	Ile/Ile	0.69	102
Head/neck cancer	GSTT1	null/deletion	null/null	0.17	632
Head/neck cancer	NAT2	481C/T = KpnI	T/T = *5/*5 = slow acetylator	0.15	105
Head/neck cancer	NAT2	590G/A = TaqI	A/A = *6/*6 = slow acetylator	0.03	105
Lung cancer	CYP1A1	Ile462Val	Val/Val	0.05	633
Lung cancer	CYP1A1	3′ C/T (MspI)	site present/site present = C/C	0.11	634
Lung cancer	CYP2E	intron 6 DraI	site present carrier	0.89	113
Lung cancer	DIA4	Ser187Pro	Ser/Ser	0.45	114
Lung cancer	GSTM1	null (gene deletion)	null/null	0.47	633
Lung cancer	MPO	−463G/A	A/A	0.08–0.09	123
Prostate cancer	AR	exon 1 GGN repeat	≤16 repeats	0.70	635
Prostate cancer	AR	exon 1 CAG repeat	<20 repeats	0.27	138
Prostate cancer	VDR	1056C/T = TaqI	C/T and T/T	0.67	146
Prostate cancer	VDR	3′ UTR poly-A; S = 14–17, L = 18–24	S/L and L/L	0.80	138
Cardiovascular disease					
CAD/MI	AGTR1	1166A/C	C	0.26–0.31	151
CAD/MI	APOA1	3′ PstI	3.3 kb allele	0.02	152
CAD/MI	APOB	Gln4154Lys = EcoRI	Lys = 13.1 kb allele	0.11	153
CAD/MI	APOB	Arg3611Glu = MspI	Glu = 9.6 kb allele	0.06	636
CAD/MI	APOB	intron 4 PvuII	Site absent	0.86	637
CAD/MI	APOB	XbaI	8.6 kb allele	0.50	153
CAD/MI	APOE	epsilon 2/3/4	epsilon 4	0.13	154
CAD/MI	ACE	intron 16 Ins/Del	Del/Del	0.24–0.29	149
CAD/MI	CYBA	His72Tyr	His/His	0.74	156
CAD/MI	F2	20210G/A	A	0.01	158
CAD/MI	F7	Arg353Gln	Arg	0.79	160
CAD/MI	GP1BA	Thr145Met = HPA2a/b	Thr/Met and Met/Met	0.15	162
CAD/MI	ITGB3	Leu33Pro = PlA1/A2	Pro = A2	0.10	167
CAD/MI	LPL	HindIII	8.7 kb homozygotes	0.34	168
CAD/MI	MTHFR	677C/T	T/T (thermolabile)	0.05–0.07	13, 14
CAD/MI	NOS3	Glu298Asp	Asp	0.07	171
CAD/MI	NOS3	intron 4 27 bp repeat	4 repeats = a allele	0.10	638
CAD/MI	PLAT	intron h Alu Ins/Del	Ins/Ins	0.30	173
CAD/MI	PON1	Arg192Gln	Arg	0.31	174
CAD/MI	SERPINE1	4G/5G in promoter	4G	0.53	181
CAD/MI	SERPINA8	Met235Thr	Thr/Thr	0.38–0.65	179, 180
DVT	F2	20210G/A	A	0.01	639
DVT	F5	Arg506Gln	Gln (Leiden)	0.02	7
DVT	MTHFR	677C/T	T/T (thermolabile)	0.18	19
HTN	ADD1	Gly460Trp	Trp	0.12–0.16	640, 641
HTN	AGTR1	1166A/C	C	0.28	195
HTN	CYP11B2	344C/T	T	0.49	196
HTN	ACE	intron 16 Ins/Del	Del/Del	0.41	193
HTN	GNB3	825C/T	T	0.25	202
HTN	NOS3	Glu298Asp	Asp	0.10–0.12	197
HTN	NOS3	intron 4 27 bp repeat	4 repeats = a allele	0.04	198
HTN	NPPA	intron 2 HpaII	Site absent	0.03	172
HTN	SERPINA8	Met235Thr	Thr	0.35–0.38	210
HTN	SERPINA8	Thr174Met	Met	0.08	210
Dermatology					
Juvenile onset psoriasis	TNF	−238G/A	A	0.04–0.05	222, 223

— Continued

Table 2
(Continued)

Disease/trait	Gene	Polymorphism	Risk allele/genotype	Frequency	Reference
Endocrinology					
Graves' disease	CTLA4	Thr17Ala	Ala	0.36	642
Male infertility	AR	CAG repeat	≥28 repeats	0.10	235
Obesity	ADRB2	Gln27Glu	Glu	0.30	238
Obesity	ADRB3	Trp64Arg	Arg	0.15	239
Osteoporosis/fracture	COL1A1	intron 1 G/T (Sp1 site)	T = s allele	0.14	250
Osteoporosis/fracture	VDR	BsmI site	B/B homozygotes	0.03	252
PCOS	CYP17	−34T/C = MspAI	C/C and C/T	0.46	254
Type 1 diabetes	CTLA4	Thr17Ala	Ala	0.33	266
Type 1 diabetes	INS	5′ VNTR	Class I allele	0.67	273
Type 1 diabetes	NEUROD1	Ala45Thr	Thr	0.05	275
Type 2 diabetes	ABCC8	exon 22 C/T(codon 761)	T	0.01–0.03	279
Type 2 diabetes	ABCC8	intron 24 -3T/C	C	0.43–0.49	279
Type 2 diabetes	GCGR	Gly40Ser	Ser	0.01–0.02	285, 286
Type 2 diabetes	GCK	3′ CA repeat	z+4 allele	0.12	287
Type 2 diabetes	GCK	5′ CA repeat	−2 allele	0.04	643
Type 2 diabetes	INS	VNTR	Class III allele = large	0.33	291
Type 2 diabetes	INSR	SstI	5.8 kb allele	0.04–0.06	292, 293
Type 2 diabetes	INSR	Val985Met	Met	0.01	644
Type 2 diabetes	IPF1	Asp76Asn	Asn	0.01	294
Type 2 diabetes	KCNJ11	Glu23Lys	Lys	0.37	296
Type 2 diabetes	PPARG	Pro12Ala	Pro	0.91	37
Type 2 diabetes	PPP1R3	Ins/Del in ARE	Del	0.48	298
Type 2 diabetes	SLC2A1	XbaI	6.2 kb = site absent	0.14–0.30	645
Type 2 diabetes	SLC2A2	TaqI	13 kb = site present	0.89	301
Type 2 diabetes	FRDA	GAA repeat	10–36 repeats	0.03–0.04	284
Type 2 diabetes	GYS1	XbaI	A2 = site present	0.04	289
Gastroenterology					
IBD	F5	Arg506Gln	Gln (Leiden)	0.06	309
Infectious disease					
HIV infection/AIDS	CCR2	Val64Ile	Val with AIDS	0.87	324
HIV infection/AIDS	CCR5	32 bp Ins/Del	Ins with infection	0.90–0.91	325, 326
Miscellaneous					
Overall mortality	APOE	epsilon 2/3/4	epsilon 4	0.22	354
Neonatal disease					
Cleft lip/palate	TGFA	TaqI	2.7 kb allele	0.05	364
Cleft lip/palate	TGFA	BamHI	4.0 kb allele	0.87	364
Neural tube defect	MTHFR	677C/T	T/T (thermolabile)	0.02–0.06	16, 17
Neural tube defect	T	intron 7 +2T/C	C	0.30	367
Neural tube defect	MTR	2756A/G	A/A and A/G	0.90	366
Neurology					
Alzheimer's disease	A2M	exon 18 5′ splice Ins/Del	Del	0.23	374
Alzheimer's disease	A2M	Val1000Ile	Val/Val	0.07	375
Alzheimer's disease	APOE	epsilon 2/3/4	epsilon 4	0.16–0.24	646, 647
Alzheimer's disease	BCHE	Ala539Thr (K variant)	Thr	0.13	648
Alzheimer's disease	BLMH	Ile443Val	Val/Val	0.07	383
Alzheimer's disease	CTSD	Ala224Val	Val	0.07	384
Alzheimer's disease	LRP1	766T/C (exon 3)	C	0.80	649
Alzheimer's disease	LRP1	tetranucleotide repeat	87 bp	0.36	650
Alzheimer's disease	SERPINA3	Ala15Thr	Ala/Ala	0.27	390
Alzheimer's disease	PSEN1	16A/C (intron 8)	A/A	0.27–0.28	389, 651
Alzheimer's disease	VLDLR	5′ UTR CGG repeat	5 repeats	0.36	396
Creutzfeldt-Jakob disease	PRNP	Met129Val	Met/Met	0.37	397
Ischemic stroke	APOE	epsilon 2/3/4	epsilon 4	0.06	652
Ischemic stroke	ACE	intron 16 Ins/Del	Del/Del	0.22	405
Ischemic stroke	F2	G20210A	A	0.01	410
Ischemic stroke	MTHFR	677C/T	T/T (thermolabile)	0.10–0.21	413, 414
Ischemic stroke	NOS3	Glu298Asp	Glu	0.61	415
Multiple sclerosis	MBP	5′ TGGA repeat	≥2.14 kb alleles	0.13	424
Parkinson's disease	COMT	Val158Met	Met/Met	0.06	653
Parkinson's disease	CYP2D6	BstMI site	B allele	0.10	435
Parkinson's disease	DRD2	intron 2 GT repeat	allele 3 = 122 bp	0.45	437
Parkinson's disease	MAOA	intron 2 GT repeat	A4 = 119 bp	0.06	440
Parkinson's disease	MAOB	intron 13 G/A	allele 1	0.45	441
Parkinson's disease	NAT2	481C/T = KpnI	T = *5 = slow acetylator	0.31	443
Parkinson's disease	SERPINA3	Ala15Thr	Ala/Ala	0.08	445
Parkinson's disease	SLC6A3	3′ UTR VNTR	11 repeats	0.01	446

— Continued

Table 2
(Continued)

Disease/trait	Gene	Polymorphism	Risk allele/genotype	Frequency	Reference
Obstetric disease					
Preeclampsia	**F5**	**Arg506Gln**	**Gln (Leiden)**	**0.02**	**459**
Preeclampsia	MTHFR	677C/T	T	0.11	461
Pharmacogenetics					
Clozapine response	**DRD3**	**Ser9Gly**	**Ser/Ser with no response**	**0.35**	**469**
Clozapine response	HTR2A	102T/C	T with no response	0.46	471
Clozapine response	HTR2A	His452Tyr	Tyr with no response	0.07	654
Drug-induced tardive dyskinesia	**DRD3**	**Ser9Gly**	**Gly/Gly with dyskinesia**	**0.04**	**478**
Tacrine Response	APOE	epsilon 2/3/4	epsilon 4 with no response	0.41	491
Psychiatry					
Anorexia	**HTR2A**	**−1438A/G**	**G**	**0.41**	**494**
ADHD	**DRD4**	**exon 3 VNTR**	**≥7 repeats**	**0.12**	**496**
Bipolar disorder	**COMT**	**Val158Met**	**Met**	**0.18**	**506**
Bipolar disorder	MAOA	CA repeat	a2 is protective	0.21	655
Bipolar disorder	MAOA	5′ VNTR	v1-v3 (long alleles)	0.61	512
Bipolar disorder	MAOA	941T/G	T	0.65	512
Bipolar disorder	**SLC6A4**	**intron 2 VNTR**	**12 repeats**	**0.54**	**517**
Bipolar disorder	TPH	intron 7 218A/C	C	0.36	518
Depression	COMT	Val158Met	Met/Val and Val/Val	0.57	522
Depression	**SLC6A4**	**intron 2 VNTR**	**9 repeats**	**0.01**	**525, 656**
Depression	SLC6A4	5′ Ins/Del (5HTTLPR)	Del/Del	0.18	657
OCD	**SLC6A4**	**5′ Ins/Del (5HTTLPR)**	**Ins**	**0.01**	**530**
Schizophrenia	APOE	epsilon 2/3/4	epsilon 4	0.15	533
Schizophrenia	COMT	Val158Met	Val	0.68	536
Schizophrenia	**DRD2**	**−141C Ins/Del**	**Ins**	**0.78**	**537**
Schizophrenia	**DRD3**	**Ser9Gly**	**Ser/Ser**	**0.37**	**538**
Schizophrenia	HMBS	intron 1 ApaLI	At least one site present	0.69	542
Schizophrenia	**KCNN3**	**second CAG repeat**	**>19 repeats**	**0.14**	**658**
Schizophrenia	**HTR2A**	**102T/C**	**C**	**0.39–0.56**	**544, 659, 660**
Schizophrenia	NTF3	5′ dinucleotide repeat	A3 = 147 bp	0.20	547
Pulmonary disease					
Asthma/atopy	**ACE**	**intron 16 Ins/Del**	**Del/Del**	**0.28**	**552**
Asthma/atopy	IL4	590C/T	T	0.70	560
Asthma/atopy	IL4R	Gln576Arg	Arg	0.10	561
Asthma/atopy	**IL4R**	**Ile50Val**	**Ile**	**0.40**	**661**
Asthma/atopy	LTA	intron 1 NcoI	5.5 kb = allele 1	0.33	563
Asthma/atopy	MS4A1	Ile181Leu	Leu/Leu and Ile/Leu	0.12	564
Asthma/atopy	**MS4A1**	**Gly237Glu**	**Glu**	**0.03**	**662**
Asthma/atopy	**TNF**	**−308G/A**	**A**	**0.18**	**563**
COPD/emphysema	**EPHX1**	**Tyr113His**	**His/His**	**0.06**	**575**
COPD/emphysema	TNF	−308G/A	A	0.02	580
COPD/emphysema	**SERPINA1**	**Taq1**	**2.4 kb allele = T2**	**0.02**	**578**
Rheumatology					
SLE	CTLA4	Thr17Ala	Ala	0.26	617
SLE	**FCGR2A**	**His131Arg**	**Arg**	**0.45–0.48**	**619**
SLE	**MBL2**	**Gly54Asp**	**Asp**	**0.09**	**623**
SLE	**TNF**	**−308G/A**	**A**	**0.11**	**663**

For each disease/trait, genes and polymorphisms within those genes are listed if there are at least three studies (and at least one achieving statistical significance) that test association between the polymorphism and the disease or trait. Gene symbols are as in Table 4. Associations with at least one replication (more than one report achieving statistical significance) are indicated in boldface. For describing the polymorphisms, standard amino acid abbreviations are used for missense polymorphisms and the start codon is numbered as 1. Where nucleotides are used to describe the polymorphism, numbering is as used in the studies and may refer to the start site of translation, transcription, or intron/exon boundary, depending on the context. Other types of polymorphisms include VNTRs (variable number tandem repeat); di-, tri-, or tetranucleotide repeats, Ins/Del (insertion deletion) polymorphisms, restriction fragment length polymorphism (indicated by the restriction enzyme used); polynucleotide tracts; or polymorphisms in the UTR (untranslated region). The allele(s) or genotype(s) conferring risk of disease are shown, and the frequency in control population(s) of the risk allele(s) or genotype(s) is indicated. The final column gives the first identified reference(s) reporting a significant association between the polymorphism and disease/trait. Full citations can be found at *www.geneticsinmedicine.org*. IBD, inflammatory bowel disease. For other abbreviations, see Table 1.

some authors believe that even minimal ethnic matching of cases and controls is adequate to prevent stratification.[36] However, there are as yet no empirical data that address the degree of stratification found in a typical association study.

LINKAGE DISEQUILIBRIUM

Failure of replication can also occur if the polymorphism being tested is not itself the causal variant but is rather in linkage disequilibrium with the causal variant. Linkage disequilibrium, in which nearby variants are correlated with each other more often than expected by chance, depends heavily on population history and on the genetic make-up of the founders of that population. If all examples of a particular stretch of DNA in a population derive from a recent common ancestor, there will have been few opportunities for recombination events to separate variants within that stretch of DNA and the variants will often be inherited together throughout the population. If, in a different population, the time since a common ancestor is

Table 3
Highly consistently reproducible associations (≥75% positive studies)

Disease/trait	Gene	Polymorphism	Risk allele/genotype	Frequency	Reference
DVT	F5	Arg506Gln	Gln (Leiden)	0.015	7
Graves' disease	CTLA4	Thr17Ala	Ala	0.62	642
Type 1 diabetes	INS	5' VNTR	Class I allele	0.67	273
HIV infection/AIDS	CCR5	32 bp Ins/Del	Del with protection	0.05–0.07	325, 326
Alzheimer's disease	APOE	epsilon 2/3/4	epsilon 4	0.16–0.24	646, 647
Creutfeldt-Jakob disease	PRNP	Met129Val	Met/Met	0.37	397

Associations between polymorphisms and disease where at least 75% of identified studies achieved statistical significance are shown; the format is as in Table 2. DVT, deep vein thrombosis.

longer, more recombination events will have occurred, disrupting linkage disequilibrium in the region. Furthermore, the particular arrangement of variants in the founders of a population will determine which variants are inherited together. Thus, it is possible that a polymorphism will be in linkage disequilibrium with a nearby disease allele in one population but not in another, leading to variable results of association studies. For example, many of the associations with *TNF* in Table 1 might reflect associations with nearby HLA loci (HLA is a region with strong linkage disequilibrium over large distances). To explore this possibility, positive associations should be followed up by testing adjacent markers (both individually and as multi-marker haplotypes). If linkage disequilibrium is present (and particularly if any of the haplotypes or adjacent markers show stronger association), the possibility exists that the original marker tested is not the causal allele, and further studies of the region are warranted. Although it should be possible to exhaustively test modest sized regions of linkage disequilibrium, special circumstances (e.g., recently admixed populations) may in theory give rise to correlation between markers at much greater distances.

GENE–GENE AND GENE–ENVIRONMENT INTERACTIONS

Another potential source of variable findings is gene–gene or gene–environment interactions that differ between populations. For example, if the effect of a variant were only manifest in populations with a particular genetic or environmental background, then association would only be seen in populations or subgroups with the appropriate genetic or environmental characteristics. This explanation is commonly invoked to explain differing results of association studies but is less frequently supported by direct evidence. A further problem arises when considering gene–gene or gene–environment interactions: when combinations of alleles and/or environmental factors are studied, *P* values are rarely corrected for the number of tests reported (much less the number of tests actually performed). Such "nominally" significant results must be considered to be the product of hypothesis generation rather than hypothesis testing and, therefore, require replication. Perhaps the best possible method of demonstrating that a gene–environment interaction is likely to be correct (and not a statistical fluctuation expected when exploring numerous hypotheses) is to divide the study population randomly into two parts and require that any findings be observed in both parts of the study. Sample sizes need to be increased slightly to maintain power, but the ability to generate and then test hypotheses in the same sample would seem to outweigh this consideration. Otherwise, one requires a replication population that is exactly matched for environmental and genetic background, an extremely unlikely scenario.

WEAK GENETIC EFFECTS AND LACK OF POWER

Finally, associations can be real but nonetheless not reproducible if the underlying genetic effect is weak. If the subsequent studies are small in size, they will be underpowered to reliably detect weak effects and, therefore, fail to achieve statistical significance. This difficulty is heightened by the "jackpot" effect, in which the first group to publish a significant association involving a weak locus is more likely to have overestimated than underestimated the true effect of the polymorphism. This phenomenon occurs because each study imprecisely estimates the strength of the effect (due to sampling variation). Because a weak effect would in most cases not provide a statistically significant finding in a typically sized study (a few hundred cases and controls), the first published study that does manage to achieve statistical significance is almost certain to have overestimated the true effect of the variant being tested. Subsequent studies thus need to include much larger numbers of patients to achieve statistical significance. In particular, failure to observe the magnitude of effect seen in the first study should not be taken as a repudiation of the association. We observed this phenomenon for the association of type 2 diabetes and a Pro12Ala polymorphism in the *PPARG* gene, where an initial study estimated the effect on diabetes risk to be threefold,[37] but subsequent studies observed very modest risks that usually did not achieve statistical significance.[38–42] We tested the variant in several large populations and found that the effect on diabetes risk was modest (1.25-fold) but significant ($P = 0.002$ in our data alone[29]). Indeed, all of the previous studies, both positive and negative, were consistent with this 1.25-fold effect, and two subsequent large studies confirmed this association.[43,44] Because many alleles may have similarly weak genetic effects, large studies and/or meta-analyses of multiple studies will often be required to determine whether genetic associations between polymorphisms and disease are significant.

Table 4
Gene symbols with OMIM numbers and aliases/descriptions

Gene symbol	OMIM #	Aliases/descriptions
A2M	103950	Alpha-2-macroglobulin
ABCC8	600509	Sulfonylurea receptor; SUR
ACE	106180	Angiotensin converting enzyme 1; DCP1; dipeptidyl carboxypeptidase 1
ACP1	171500	Acid phosphatase 1, soluble (erythrocyte)
ADA	102700	Adenosine deaminase
ADD1	102680	Alpha adducin 1
ADH1B	103720	ADH2; class I alcohol dehydrogenase, beta polypeptide
ADH4	103740	Alcohol dehydrogenase 4
ADORA2A	102776	Adenosine A2a receptor; ADORA2; RDC8
ADPRT	173870	ADP-ribosyltransferase; poly(ADP) ribose polymerase; PARP
ADRB2	109690	Beta 2 adrenergic receptor
ADRB3	109691	Beta 3 adrenergic receptor
AGTR1	106165	Angiotensin II receptor, type 1
ALAD	125270	Delta-aminolevulinate dehydratase
ALDH2	100650	Aldehyde dehydrogenase 2
ALOX5	152390	Arachidonate 5-lipoxygenase
AMPD1	102770	Adenosine monophosphate deaminase 1; MADA
APBB1	602709	amyloid beta precursor protein-binding, family B, member 1; FE65
APC	175100	Adenomatosis polyposis coli
APOA1	107680	Apolipoprotein A-I
APOA4	107690	Apolipoprotein A-IV
APOB	107730	Apolipoprotein B
APOC1	107710	Apolipoprotein C-I
APOC2	207750	Apolipoprotein C-II
APOD	107740	Apolipoprotein D
APOE	107741	Apolipoprotein E
AR	313700	Androgen receptor
ATP1A3	182350	ATPase, Na+/K+ transporting, alpha 3 polypeptide
BCHE	177400	Butyrylcholinesterase
BCL2	151430	B-cell CLL/lymphoma 2
BCL3	109560	B-cell leukemia/lymphoma 3
BDKRB1	600337	Bradykinin B1 receptor; kinin B1 receptor
BLMH	602403	Bleomycin hydrolase
C4A	120810	C4; complement component 4A
C4B	120820	Complement component 4B; C4F
CCK	118440	Cholecystokinin
CCKBR	118445	Cholecystokinin B receptor; gastrin receptor
CCR2	610267	Chemokine (C-C motif) receptor 2; CKR2; CMKBR2
CCR5	601373	Chemokine (C-C motif) receptor 5; CKR5; CMKBR5
CD14	158120	CD14 antigen
CD36	173510	Thrombospondin receptor; collagen type I receptor; fatty acid translocase; platelet glycoprotein IIIb; GPIIIb
CD3D	186790	CD3, delta subunit; T3D; TiT3 complex
CD4	186940	CD4 antigen (p55); T4/LEU3
CDKN1A	116899	Cyclin-dependent kinase inhibitor 1A; p21; Cip1; WAF1
CDSN	602593	S gene (corneodesmosin)
CETP	118470	Cholesteryl ester transfer protein
CFTR	602421	Cystic fibrosis transmembrane conductance regulator; ATP-binding cassette (sub-family C, member 7); ABCC7
CHRNA4	118504	Neuronal nicotinic acetylcholine receptor, alpha-4 subunit
CMA1	118938	Mast cell chymase 1
COL1A1	120150	Collagen type I alpha 1
COL2A1	120140	Collagen, type II, alpha 1; chondrocalcin; COL11A3
COL9A2	600204	Collagen, type IX, alpha 2; EDM2
COMT	116790	Catechol O-methyltransferase
CRH	122560	Corticotropin releasing hormone
CTLA4	123890	Cytotoxic T-lymphocyte-associated protein 4; CD152
CTSD	116840	Cathepsin D; lysosomal aspartyl protease
CX3CR1	601470	Chemokine (C-X3-C) receptor 1
CYBA	233690	Cytochrome b-245 alpha; p22-PHOX
CYP11A	118485	Cholesterol side chain cleavage enzyme; cytochrome P450, subfamily XIA
CYP11B2	124080	Aldosterone synthase; steroid 11-beta-hydroxylase; cytochrome P450, subfamily XIB, polypeptide 2
CYP17	202110	17-alpha-hydroxylase; 17,20 lyase; cytochrome P450, subfamily XVII
CYP19	107910	Aromatase; cytochrome P450, subfamily XIX
CYP1A1	108330	Cytochrome P450, subfamily IA, polypeptide 1
CYP1B1	601771	Cytochrome P450, subfamily IB, polypeptide 1
CYP2A6	122720	Cytochrome P450, subfamily IIA, polypeptide 6; coumarin 7-hydroxylase
CYP2C19	124020	Cytochrome P450, subfamily IIC, polypeptide 19
CYP2C9	601130	Cytochrome P450, subfamily IIC, polypeptide 9
CYP2D6	124030	Cytochrome P450, subfamily IID, polypeptide 6; debrisoquine 4-hydroxylase
CYP2E	124040	Cytochrome P450, subfamily IIE; CYP2E1
CYP3A4	124010	Cytochrome P450, subfamily 3A, polypeptide 4; glucocorticoid-inducible P450

—Continued

Table 4
(Continued)

Gene symbol	OMIM #	Aliases/descriptions
DBH	223360	Dopamine beta-hydroxylase
DDC	107930	Dopa decarboxylase; aromatic L-amino acid decarboxylase
DIA4	125860	NQO1; Diaphorase; NAD(P)H:quinone oxidoreductase
DLST	126063	Alphaketoglutarate dehydrogenase, E2 subunit; dihydrolipoamide S-succinyltransferase
DRD1	126449	Dopamine receptor D1
DRD2	126450	Dopamine receptor D2
DRD3	126451	Dopamine receptor D3
DRD4	126452	Dopamine receptor D4
DRD5	126453	Dopamine receptor D5
EDNRA	131243	Endothelin receptor type A
ELAC2	605367	HPC2; elaC (E. coli) homolog 2; prostate cancer, hereditary, 2
EN2	131310	Engrailed homolog 2
ENG	131195	Endoglin; CD105; Osler-Rendu-Weber syndrome 1
EPHX1	132810	Microsomal epoxide hydrolase 1 (xenobiotic)
ERBB2	164870	HER-2; NEU; v-erb-b2 avian erythroblastic leukemia viral oncogene homolog 2; NGL
ERCC1	126380	Excision repair cross-complementing rodent repair deficiency, complementation group 1
ESR1	133430	Estrogen receptor 1; estrogen receptor alpha
ETS1	164720	v-ets avian erythroblastosis virus E26 oncogene homolog 1
F13A1	134570	Coagulation factor XIII, A1 polypeptide
F2	176930	Prothrombin; coagulation factor II
F3	134390	Tissue factor; thromboplastin; coagulation factor III
F5	227400	Coagulation factor V
F7	227500	Coagulation factor VII
FCGR2A	146790	CD32; Fc IgG low affinity IIa receptor
FCGR3A	146740	CD16; Fc frament of IgG, low affinity receptor IIIa
FGA	134820	Fibrinogen alpha, A polypeptide
FGB	134830	Fibrinogen, B beta polypeptide
FMR1	309550	Fragile X mental retardation 1; FRAXA
FRDA	229300	Frataxin; Friedreich ataxia; X25
FSHB	136530	Follicle stimulation hormone, beta polypeptide
FST	136470	Follistatin
GABRA5	137142	Gamma-aminobutyric acid (GABA) receptor A, alpha 5
GABRB3	137192	Gamma-aminobutyric acid (GABA) receptor A, beta 3
GC	139200	Vitamin D binding protein; group-specific component
GCGR	138033	Glucagon receptor
GCK	138079	Glucokinase; MODY2; hexokinase 4
GNAL	139312	G(olf) alpha; G protein, alpha activating activity polypeptide, olfactory type
GNAS1	139320	G-protein alpha stimulating activity polypeptide 1
GNB3	139130	G-protein beta, polypeptide 3
GP1BA	231200	Platelet glycoprotein Ib, alpha polypeptide
GPX1	138320	Glutathione peroxidase
GRIK1	138245	Glutamate receptor, ionotropic, kainate 1; glutamate receptor 5
GSTM1	138350	Glutathione S-transferase M1; glutathione S-transferase mu-1
GSTM3	138390	Glutathione S-transferase M3 (brain)
GSTP1	134660	Glutathione S-transferase pi; GST3
GSTT1	600436	Glutathione S-transferase theta 1
GYS1	138570	Glycogen synthase (muscle); GYS
HFE	235200	Hemochromatosis; HLAH
HMBS	176000	Porphobilinogen deaminase; hydroxymethylbilane synthase; PBGD
HNMT	605238	Histamine N-methyltransferase
HRAS	190020	v-Ha-ras Harvey rat sarcoma viral oncogene homolog; HRAS1
HRH2	142703	Histamine receptor H2
HSD11B2	218030	11-beta hydroxysteroid dehydrogenase 2; AME
HSPA1A	140550	Heat shock 70kD protein 1A; hsp70-1
HSPA2	140560	Heat shock 70kD protein 2; hsp70-2
HSPA8	600816	Heat shock 70kD protein 8; HSC70
HTR1B	182131	5-hydroxytryptamine (serotonin) receptor 1B; 5HT1D(beta)
HTR2A	182135	5-hydroxytryptamine (serotonin) receptor 2A; HTR2
HTR2C	312861	5-hydroxytryptamine (serotonin) receptor 2C; HTR1C
HTR5A	601305	5-hydroxytryptamine (serotonin) receptor 5A
HTR6	601109	5-hydroxytryptamine (serotonin) receptor 6
ICAM1	147840	Intercellular adhesion molecule 1; CD54
IFNG	147570	Interferon gamma
IGF2	147470	Insulin-like growth factor II; somatomedin A
IGHV2-5	600949	Immunoglobulin heavy chain variable region V2-B5
IGHV3-30-5	147070	Humhv3005; immunoglobulin heavy chain variable region
IL10	124092	Interleukin 10
IL13	147683	Interleukin 13
IL1A	147760	Interleukin 1-alpha
IL1B	147720	Interleukin 1-beta

— Continued

Table 4
(Continued)

Gene symbol	OMIM #	Aliases/descriptions
IL1RN	147679	Interleukin 1 receptor antagonist; IL1RA
IL4	147780	Interleukin 4; BSF1
IL4R	147781	Interleukin 4 receptor
IL6	147620	Interleukin 6; interferon, beta 2; B-cell differentiation factor; BSF2; HSF
IL8	146930	Interleukin 8; NAP1; SCYB8; monocyte-derived neutrophil chemotactic factor
IL9R	300007	Interleukin 9 receptor
IMPA1	602064	Inositol(myo)-1(or 4)-monophosphatase 1
INS	176730	Insulin
INSR	147670	Insulin receptor
IPF1	600733	Insulin promoter factor 1; PDX1; IDX1; STF1; MODY4
IRS1	147545	Insulin receptor substrate 1
ITGA2	192974	Platelet glycoprotein Ia/IIa; integrin, alpha-2 ; CD49B; VLA2 receptor, alpha-2 subunit
ITGB3	173470	Glycoprotein IIIa; integrin, beta-3; CD61
KCNJ11	600937	Kir6.2; BIR; potassium inwardly-rectifying channel, subfamily J, member 11
KCNN3	602983	hKCa3; SKCA3; SK3; hSK3; potassium intermediate/small conductance calcium-activated channel, subfamily N, member 3
KLKB1	229000	Kallikrein B, plasma; formerly KLK3
LDLR	143890	Low density lipoprotein receptor; familial hypercholesterolemia
LEP	164160	Leptin; Ob
LHB	152780	Luteinizing hormone, beta polypeptide
LIPE	151750	Hormone sensitive lipase
LPL	238600	Lipoprotein lipase
LRP1	107770	Low density lipoprotein-related protein 1; alpha-2-macroglobulin receptor; ApoE receptor
LTA	153440	TNF beta; lymphotoxin A; TNF superfamily, member 1
MAOA	309850	Monoamine oxidase A
MAOB	309860	Monoamine oxidase B; MAO, platelet; MAO, brain
MAPT	157140	Microtubule-associated protein tau; MTBT1
MBL2	154545	Mannose binding lectin; mannose binding protein; MBP1
MBP	159430	Myelin basic protein
MC1R	155555	Melanocortin 1 receptor; alpha melanocyte stimulating hormone receptor; MSHR
MGMT	156569	O-6-methylguanine-DNA methyltransferase
MLH1	120436	MutL (E. coli) homolog 1; colon cancer, nonpolyposis type 2; HNPCC2
MMP1	120353	Matrix metalloproteinase 1; interstitial collagenase
MMP3	185250	Matrix metalloproteinase 3; stromelysin 1; progelatinase
MMP9	120361	Matrix metalloproteinase 9; gelatinase B; 92kD type IV collagenase
MPO	254600	Myeloperoxidase
MS4A1	147138	Fc IgE receptor; Membrane-spanning 4-domains, subfamily A, member 1
MSH3	600887	MutS (E. coli) homolog 3
MSX1	142983	Msh (Drosophila) homeo box homolog 1; HOX7; HYD1
MTHFR	236250	5,10-methylene tetrahydrofolate reductase
MTR	156570	Methionine synthase; 5-methyltetrahydrofolate-homocysteine methyltransferase
MUC1	158340	Mucin 1, transmembrane
MUC3A	158371	Mucin 3A, intestinal; MUC3
MYC	190080	v-myc avian myelocytomatosis viral oncogene homolog
MYCL1	164850	L-myc; v-myc avian myelocytomatosis viral oncogene homolog 1, lung carcinoma derived
NAT1	108345	N-acetyltransferase 1; arylamine N-acetyltransferase 1; AAC1
NAT2	243400	N-acetyltransferase 2; arylamine N-acetyltransferase 2; AAC2
NEUROD1	601724	Neurogenic differentiation; beta2
NMB	162340	Neuromedin B
NOS1	163731	Neuronal nitric oxide synthase
NOS2A	163730	Inducible nitric oxide synthase
NOS3	163729	Endothelial nitric oxide synthase; ENOS
NPPA	108780	Natriuretic peptide precursor A; atrial natriuretic polypeptide; ANP; ANF
NPY5R	602001	Neuropeptide Y receptor Y5
NTF3	162660	Neurotrophin 3; neurotrophic factor 3; NT3
OPRM1	600018	Opioid receptor, mu 1
OPRS1	601978	Type 1 sigma receptor; SR-BP1; sigma receptor (SR31747 binding protein 1)
PCSK2	162151	Prohormone convertase 2; proprotein convertase subtilisin/kexin type 2; PC2
PGR	264080	Progesterone receptor; PR
PLA2G1B	172410	Phospholipase A2, group IB (pancreas); pancreatic phospholipase; PLA2A; PLA2
PLA2G4A	600522	Phospholipase A2, group IVA (cytosolic); cPLA2
PLA2G7	601690	Platelet-activating factor acetylhydrolase; phospholipase A2 group VII
PLAT	173370	TPA; tissue plasminogen activator
PLCG1	172420	Phospholipase C, gamma 1; PLC1; phospholipase C-148; PLC148
PON1	168820	Paraoxonase 1
PON2	602447	Paraoxonase 2
PPARG	601487	Peroxisome proliferator-activated receptor, gamma; PPAR gamma
PPP1R3	600917	Protein phosphatase 1, regulatory (inhibitor) subunit 3
PRNP	176640	Prion protein; PRP
PRTN3	177020	Proteinase 3 (serine proteinase, neutrophil, Wegener's granulomatosis autoantigen); AGP7; p29

— Continued

Table 4
(Continued)

Gene symbol	OMIM #	Aliases/descriptions
PSEN1	104311	Presenilin 1; PS1; AD3
PSMB8	177046	Proteasome subunit beta type 8; LMP7; large multifunctional protease 7
PTPRC	151460	Protein tyrosine phosphatase, receptor type, C; CD45; Ly5 homolog
RARA	180240	Retinoic acid receptor, alpha
REN	179820	Renin
RRAD	179503	RAD1; RAD; ras-related associated with diabetes
SAH	145505	SA homolog; SA
SCNN1B	600760	Epithelial sodium channel, beta subunit; ENaCb; sodium channel, non-voltage gated 1, beta
SCYA5	187011	Small inducible cytokine A5 (RANTES)
SDF1	600835	Stromal cell-derived factor 1; CXCL12
SELE	131210	E selectin; ELAM1; endothelial adhesion molecule 1; CD 62E
SELP	173610	P-selectin; PSEL; CD62 antigen; CD62P; platelet alpha granule membrane protein 140kD; GRMP
SERPINA1	107400	Alpha-1-antitrypsin; protease inhibitor 1; PI
SERPINA3	107280	Alpha-1-antichymotrypsin; AACT
SERPINA8	106150	Angiotensinogen; AGT
SERPINE1	173360	Plasminogen activator inhibitor 1; PAI1
SFTPA1	178630	Pulmonary surfactant apoprotein; SPA; SP-A
SHBG	182205	Sex hormone-binding globulin
SLC11A1	600266	NRAMP1; natural resistance-associated macrophage protein 1
SLC2A1	138140	GLUT1; glucose transporter 1
SLC2A2	138160	GLUT2; glucose transporter 2
SLC6A3	126455	Dopamine transporter; DAT1
SLC6A4	182138	Serotonin transporter; 5HTT; SERT
SNAP25	600322	SNAP-25; synaptosomal-associated protein, 25 kDa
SNCA	163890	Synuclein, alpha
SOD2	147460	Superoxide dismutase 2, mitochondrial; manganese superoxide dismutase; MnSOD
SRD5A2	264600	Steroid-5-alpha-reductase, alpha polypeptide 2
T	601397	T Brachyury (mouse) homolog
TAP1	170260	Antigen peptide transporter; ABCB2
TAP2	170261	Antigen peptide transporter 2; ATP-binding cassette, sub-family B, member 2; ABCB2
TBXA2R	188070	Thromboxane A2 receptor
TCF1	142410	HNF1-alpha; transcription factor 1, hepatic
TF	190000	Transferrin
TFCP2	189889	Transcription factor CP2
TGFA	190170	Transforming growth factor, alpha
TGFB1	190180	Transforming growth factor, beta
TGFB3	190230	Transforming growth factor beta 3
TH	191290	Tyrosine hydroxylase
THBD	188040	Thrombomodulin; THRM; CD141
THRB	190160	Thyroid hormone receptor beta; ERBA2
TNF	191160	Tumor necrosis factor alpha; TNFA; TNF superfamily, member 2
TNFRSF6	134637	Fas antigen; CD95; tumor necrosis receptor superfamily, member 6; APT1
TP53	191170	Tumor protein p53
TPH	191060	Tryptophan hydroxylase
TPMT	187680	Thiopurine S-methyltransferase
TRA@	186880	T-cell receptor alpha locus
TRD@	186810	T-cell receptor delta locus
TRHR	188545	Thyrotropin-releasing hormone receptor
UCHL1	191342	Ubiquitin carboxy-terminal hydrolase L1; ubiquitin carboxy-terminal esterase L1
UGB	192020	CC16; uteroglobin; Clara cell-specific 16-kDa protein; CC10; CCSP
UGT1A1	191740	UDP glycosyltransferase 1 family, polypeptide A1; UDP-glucuronosyltransferase phenol/bilirubin; UGT1A
VDR	601769	Vitamin D receptor
VLDLR	192977	Very low density lipoprotein receptor
WFS1	222300	Wolfram syndrome 1; wolframin
WRN	604611	Werner syndrome; DNA helicase, recQ-like, type 3; RECQ3; RECQL2
XRCC3	600675	X-ray repair cross-complementing protein 3
YWHAH	113508	14-3-3 eta; tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, eta polypeptide

For each gene, the official gene symbol from the Human Genome Organisation (HUGO, *http://www.gene.ucl.ac.uk/nomenclature*), the number for the Online Mendelian inheritance in man (OMIM, Baltimore: Johns Hopkins University, Center for Medical Genetics, 1996, *http://www3.ncbi.nlm.nih.gov/omim*), and common aliases and descriptions are given.

GENERAL CONCLUSIONS

How does one tell whether reported associations between polymorphisms and disease are real? Reasonable criteria for declaring association have been proposed, including low *P* values, replication in multiple samples, and avoidance of population stratification (such as by using family-based controls[28]). However, most studies do not meet these criteria, and multiple studies of an association are usually inconsistent. In these cases, meta-analysis of all published studies may guide interpretation, and we strongly advocate that any publication of an association study (whether negative or positive) be accompanied by a meta-analysis of all similar studies. Accordingly, in-

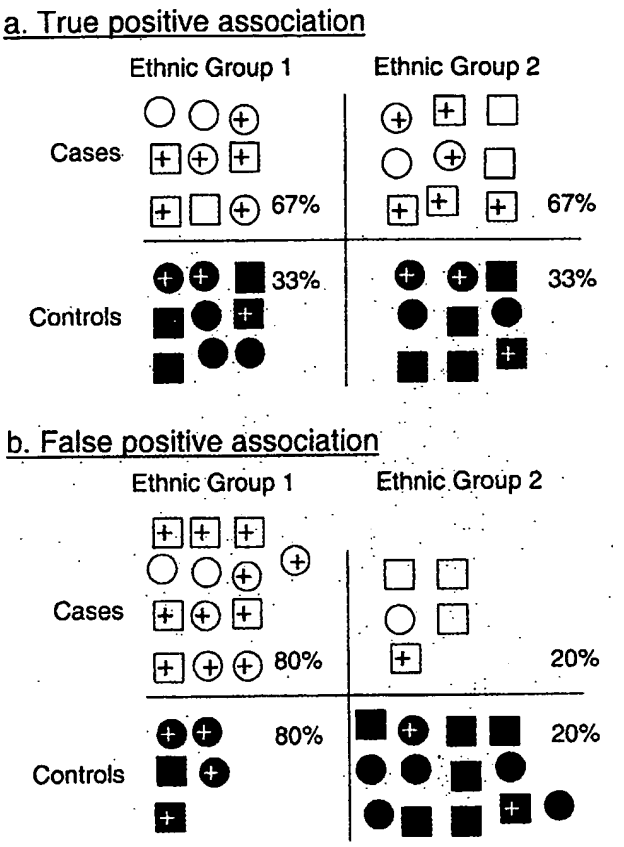


Fig. 2 True associations contrasted with false-positive associations due to ethnic admixture. The open shapes represent individuals with disease, and the filled shapes represent individuals from a control population. Shapes with a plus sign (+) represent individuals carrying the putative risk allele being tested for association. In both figures, the fraction of individuals carrying the risk allele is twice as large in the case population as in the control population. Figure 2a (Top): True-positive association: the frequency of the risk allele is greater in cases than in controls in both ethnic groups. Figure 2b (Bottom): False-positive association due to ethnic admixture: the frequency of the risk allele is identical in cases and controls in both populations. However, the allele is twice as frequent overall in cases as in controls. This false appearance of association is due to ethnic admixture, i.e., ethnic group 1 is overrepresented in the cases, and the allele being tested is prevalent in ethnic group 1 but not ethnic group 2.

dividual researchers should also publish or make easily available sufficient information to facilitate future meta-analysis, including relevant genotype and phenotype data. Publication bias may present a major challenge to such analyses, because the omission of small negative studies will bias the pooled data toward a positive result. In this regard, we advocate a mechanism for storage and dissemination of all association data (published or not), perhaps in a widely accepted and curated Web site and/or in brief "negative results" sections of specialty journals. Until complete meta-analyses can be performed using data from multiple large studies, we will be left with a scenario in which the majority of reported associations are in genetic purgatory, neither convincingly confirmed or refuted, awaiting future judgment.

Much of the interest surrounding genetic association studies centers on the potential clinical application of polymorphisms that serve as markers for disease. In particular, it has been proposed that these markers can both serve as predictors of disease and as a means to tailor treatment of disease. Although this scenario may well become reality, the current irreproducibility of most studies should raise a loud cautionary alarm. Certainly, clinical applications of genetic associations should not be considered until the degree of certainty far exceeds the level currently achieved for the vast majority of such associations. Furthermore, even if an association is supported by extremely convincing evidence, screening patients is only appropriate if determining an individual's genotype would allow a clinically proven beneficial intervention that outweighs the risk of performing the test. Genetic tests also give rise to ethical considerations, because of the implication for family members, the potential for discrimination, the immutability of genetic risk factors, and the predictive nature of such tests. (Although, given the probable modest effects of any particular genetic variant, most genetic tests are likely to be much less predictive of future health than widely used screens such as blood pressure and cholesterol measurements.) Societal consensus and legislative solutions addressing these ethical concerns are needed before such testing enters widespread clinical practice.

Because of the scientific and ethical uncertainties, a "DNA chip" that can determine crucial genotypes and accurately predict future health is unlikely to become a widespread and useful screening tool in the near future, even if concerns regarding reproducibility can be resolved. Rather, the most likely short-term benefit from genetic association studies will be a better understanding of disease pathogenesis, which will hopefully lead in turn to novel and better treatments and/or more tailored drug therapy. If genetic association studies can provide these sorts of advances, they will have proven a valuable resource in the struggle to understand and treat common disease.

Acknowledgments

J.N.H. is a recipient of a Postdoctoral Fellowship for Physicians from the Howard Hughes Medical Institute and a Burroughs Wellcome Career Award in the Biomedical Sciences. K.L and E.B. were supported by grants from Affymetrix Inc., Millennium Pharmaceuticals, Inc., and Bristol-Myers Squibb Company to Eric S. Lander, Whitehead/MIT Center for Genome Research, Cambridge, Massachusetts. We thank David Altshuler, Pamela Sklar, Eric Lander, and C. Leigh Pearce for helpful comments, and Delores Gray for assistance in locating manuscripts. Supplementary information (full citations for references 45 to 663 and Supplementary Table 1) can be found at *www.geneticsinmedicine.org*.

References

1. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science* 1994;265:2037–2048.
2. Chakravarti A. Population genetics–making sense out of sequence. *Nat Genet* 1999; 21:56–60.
3. Risch NJ. Searching for genetic determinants in the new millennium. *Nature* 2000; 405:847–856.
4. Harris H. The principles of human biochemical genetics. Amsterdam: North-Holland Publishing Company, 1975.

5. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES, Atshuler D. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 2001;409:928–933.
6. Dahlback B, Carlsson M, Svensson PJ. Familial thrombophilia due to a previously unrecognized mechanism characterized by poor anticoagulant response to activated protein C: prediction of a cofactor to activated protein C. *Proc Natl Acad Sci U S A* 1993;90:1004–1008.
7. Bertina RM, Koeleman BP, Koster T, Rosendaal FR, Dirven RJ, de Ronde H, van der Velden PA, Reitsma PH. Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* 1994;369:64–67.
8. Aparicio C, Dahlback B. Molecular mechanisms of activated protein C resistance. Properties of factor V isolated from an individual with homozygosity for the Arg506 to Gln mutation in the factor V gene. *Biochem J* 1996;313:467–472.
9. Salomon O, Steinberg DM, Zivelin A, Gitel S, Dardik R, Rosenberg N, Berliner S, Inbal A, Many A, Lubetsky A, Varon D, Martinowitz U, Seligsohn U. Single and combined prothrombotic factors in patients with idiopathic venous thromboembolism: prevalence and risk assessment. *Arterioscler Thromb Vasc Biol* 1999;19:511–518.
10. Dilley A, Austin H, Hooper WC, El-Jamil M, Whitsett C, Wenger NK, Benson J, Evatt B. Prevalence of the prothrombin 20210 G-to-A variant in blacks: infants, patients with venous thrombosis, patients with myocardial infarction, and control subjects. *J Lab Clin Med* 1998;132:452–455.
11. Kang SS, Wong PW, Susmano A, Sora J, Norusis M, Ruggie N. Thermolabile methylenetetrahydrofolate reductase: an inherited risk factor for coronary artery disease. *Am J Hum Genet* 1991;48:536–545.
12. Frosst P, Blom HJ, Milos R, Goyette P, Sheppard CA, Matthews RG, Boers GJ, den Heijer M, Kluijtmans LA, van den Heuvel LP, et al. A candidate genetic risk factor for vascular disease: a common mutation in methylenetetrahydrofolate reductase. *Nat Genet* 1995;10:111–113.
13. Gallagher PM, Meleady R, Shields DC, Tan KS, McMaster D, Rozen R, Evans A, Graham IM, Whitehead AS. Homocysteine and risk of premature coronary heart disease. Evidence for a common gene mutation. *Circulation* 1996;94:2154–2158.
14. Kluijtmans LA, van den Heuvel LP, Boers GH, Frosst P, Stevens EM, van Oost BA, den Heijer M, Trijbels FJ, Rozen R, Blom HJ. Molecular genetic analysis in mild hyperhomocystinemia: a common mutation in the methylenetetrahydrofolate reductase gene is a genetic risk factor for cardiovascular disease. *Am J Hum Genet* 1996;58:35–41.
15. Bailey LB, Gregory JF III. Polymorphisms of methylenetetrahydrofolate reductase and other enzymes: metabolic significance, risks and impact on folate requirement. *J Nutr* 1999;129:919–922.
16. van der Put NM, Steegers-Theunissen RP, Frosst P, Trijbels FJ, Eskes TK, van den Heuvel LP, Mariman EC, den Heyer M, Rozen R, Blom HJ. Mutated methylenetetrahydrofolate reductase as a risk factor for spina bifida. *Lancet* 1995;346:1070–1071.
17. Whitehead AS, Gallagher P, Mills JL, Kirke PN, Burke H, Molloy AM, Weir DG, Shields DC, Scott JM. A genetic defect in 5,10 methylenetetrahydrofolate reductase in neural tube defects. *Q J Med* 1995;88:763–766.
18. Ma J, Stampfer MJ, Giovannucci E, Artigas C, Hunter DJ, Fuchs C, Willett WC, Selhub J, Hennekens CH, Rozen R. Methylenetetrahydrofolate reductase polymorphism, dietary interactions, and risk of colorectal cancer. *Cancer Res* 1997;57:1098–1102.
19. Margaglione M, D'Andrea G, d'Addedda M, Giuliani N, Cappucci G, Iannaccone L, Vecchione G, Grandone E, Brancaccio V, Di Minno G. The methylenetetrahydrofolate reductase TT677 genotype is associated with venous thrombosis independently of the coexistence of the FV Leiden and the prothrombin A20210 mutation. *Thromb Haemost* 1998;79:907–911.
20. Skibola CF, Smith MT, Kane E, Roman E, Rollinson S, Cartwright RA, Morgan G. Polymorphisms in the methylenetetrahydrofolate reductase gene are associated with susceptibility to acute leukemia in adults. *Proc Natl Acad Sci U S A* 1999;96:12810–12815.
21. Ma J, Stampfer MJ, Hennekens CH, Frosst P, Selhub J, Horsford J, Malinow MR, Willett WC, Rozen R. Methylenetetrahydrofolate reductase polymorphism, plasma folate, homocysteine, and risk of myocardial infarction in US physicians. *Circulation* 1996;94:2410–2416.
22. Kluijtmans LA, den Heijer M, Reitsma PH, Heil SG, Blom HJ, Rosendaal FR. Thermolabile methylenetetrahydrofolate reductase and factor V Leiden in the risk of deep-vein thrombosis. *Thromb Haemost* 1998;79:254–258.
23. Morrison K, Papapetrou C, Hol FA, Mariman EC, Lynch SA, Burn J, Edwards YH. Susceptibility to spina bifida: an association study of five candidate genes. *Ann Hum Genet* 1998;62:379–396.
24. Osuntokun BO, Sahota A, Ogunniyi AO, Gureje O, Baiyewu O, Adeyinka A, Oluwole SO, Komolafe O, Hall KS, Unverzagt FW, et al. Lack of an association between apolipoprotein E epsilon 4 and Alzheimer's disease in elderly Nigerians. *Ann Neurol* 1995;38:463–465.
25. Farrer LA, Cupples LA, Haines JL, Hyman B, Kukull WA, Mayeux R, Myers RH, Pericak-Vance MA, Risch N, van Duijn CM. Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA* 1997;278:1349–1356.
26. Tang MX, Stern Y, Marder K, Bell K, Gurland B, Lantigua R, Andrews H, Feng L, Tycko B, Mayeux R. The APOE-epsilon4 allele and the risk of Alzheimer disease among African Americans, whites, and Hispanics. *JAMA* 1998;279:751–755.
27. Altshuler D, Kruglyak L, Lander E. Genetic polymorphisms and disease. *N Engl J Med* 1998;338:1626.
28. Freely associating. *Nat Genet* 1999;22:1–2.
29. Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, Tuomi T, Gaudet D, Hudson TJ, Daly M, Groop L, Lander ES. The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* 2000;26:76–80.
30. Cardon LR, Bell JI. Association study designs for complex diseases. *Nat Rev Genet* 2001;2:91–99.
31. Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 1993;52:506–516.
32. Pritchard JK, Rosenberg NA. Use of unlinked genetic markers to detect population stratification in association studies. *Am J Hum Genet* 1999;65:220–228.
33. Devlin B, Roeder K. Genomic control for association studies. *Biometrics* 1999;55:997–1004.
34. Reich DE, Goldstein DB. Detecting association in a case-control study while correcting for population stratification. *Genet Epidemiol* 2001;20:4–16.
35. Pritchard JK, Stephens M, Rosenberg NA, Donnelly P. Association mapping in structured populations. *Am J Hum Genet* 2000;67:170–181.
36. Morton NE, Collins A. Tests and estimates of allelic association in complex inheritance. *Proc Natl Acad Sci U S A* 1998;95:11389–11393.
37. Deeb SS, Fajas L, Nemoto M, Pihlajamaki J, Mykkanen L, Kuusisto J, Laakso M, Fujimoto W, Auwerx J. A Pro12Ala substitution in PPARgamma2 associated with decreased receptor activity, lower body mass index and improved insulin sensitivity. *Nat Genet* 1998;20:284–287.
38. Mancini FP, Vaccaro O, Sabatino L, Tufano A, Rivellese AA, Riccardi G, Colantuoni V. Pro12Ala substitution in the peroxisome proliferator-activated receptor- gamma2 is not associated with type 2 diabetes. *Diabetes* 1999;48:1466–1468.
39. Ringel J, Engeli S, Distler A, Sharma AM. Pro12Ala missense mutation of the peroxisome proliferator activated receptor gamma and diabetes mellitus. *Biochem Biophys Res Commun* 1999;254:450–453.
40. Clement K, Hercberg S, Passinge B, Galan P, Varroud-Vial M, Shuldiner AR, Beamer BA, Charpentier G, Guy-Grand B, Froguel P, Vaisse C. The Pro115Gln and Pro12Ala PPAR gamma gene mutations in obesity and type 2 diabetes. *Int J Obes Relat Metab Disord* 2000;24:391–393.
41. Hara K, Okada T, Tobe K, Yasuda K, Mori Y, Kadowaki H, Hagura R, Akanuma Y, Kimura S, Ito C, Kadowaki T. The Pro12Ala polymorphism in PPAR gamma2 may confer resistance to type 2 diabetes. *Biochem Biophys Res Commun* 2000;271:212–216.
42. Meirhaeghe A, Fajas L, Helbecque N, Cottel D, Auwerx J, Deeb SS, Amouyel P. Impact of the peroxisome proliferator activated receptor gamma2 Pro12Ala polymorphism on adiposity, lipids and non-insulin-dependent diabetes mellitus. *Int J Obes Relat Metab Disord* 2000;24:195–199.
43. Douglas JA, Erdos MR, Watanabe RM, Braun A, Johnston CL, Oeth P, Mohlke KL, Valle TT, Ehnholm C, Buchanan TA, Bergman RN, Collins FS, Boehnke M, Tuomilehto J. The peroxisome proliferator-activated receptor-gamma2 Pro12A1a variant: association with type 2 diabetes and trait differences. *Diabetes* 2001;50:886–890.
44. Mori H, Ikegami H, Kawaguchi Y, Seino S, Yokoi N, Takeda J, Inoue I, Seino Y, Yasuda K, Hanafusa T, Yamagata K, Awata T, Kadowaki T, Hara K, Yamada N, Gotoda T, Iwasaki N, Iwamoto Y, Sanke T, Nanjo K, Oka Y, Matsutani A, Maeda E, Kasuga M. The Pro12 -> Ala substitution in PPAR-gamma is associated with resistance to development of diabetes in the general population: possible involvement in impairment of insulin secretion in individuals with type 2 diabetes. *Diabetes* 2001;50:891–894.